Data Mining Driven Visual Pattern Discovery

RBEI-IYER-MC2

Manik Singhal, Prakash Lekkala, Shivashankar M R, Parameshwaran S Iyer
G1/PJ-DM, Robert Bosch Engineering & Business Solutions Ltd
Bangalore, India
parameshwaran.iyer@bosch.com

***Abstract* — In the following work we present a short description of data mining methods & techniques we apply to create a visual framework for VAST Mini Challenge 2 (VAST-MC2). We provide a brief description of the problem, followed by our approach to arriving on a visual framework. From our framework, post visual mining for patterns; we list some of our key findings.**

***Index Terms* — Spatial Temporal clustering, Markov Cluster Algorithm, Event-Information Framework Ripley-L factor.**

I. Introduction

In the VAST MC2, participants have been provided with spatial temporal data related to missing employees of GAS-tech. Participants have been asked to analyze the data to determine any illegal activities, and suggest suspect employees and their locations.

We propose both a delayed, as well as a real time visual framework to identify unusual patterns. In both applications we use outputs from data mining methods as the fundamental building blocks for the visualization layer.

For the analysis conducted in hindsight, given 2 weeks of data we first look for spatial temporal clusters of employees. Given these associations between employees occur in time, we next check for locations or hotspots in space where these occur. Next we check for patterns & consistency of these events occurring over the period of the data. In real time we propose to use the Ripley-L factor [1, 2] for tracking spatial homogeneity over time. Given instantaneous locations of employees, we use these coordinate positions in computing the L-factor score. Low values of L-factor indicate high levels of spatial cohesion occurring at a given time point.

II. Assumptions & Handling Data Uncertainties

Meta data of GPS logs, consisting of car assignments, names & profiles within the organization, of 35 assigned car users of GAS-Tech along with the separate logs for 5 company trucks were made available, along with transaction logs (credit & loyalty transactions) of 54 GAS-Tech employees were also provided. These 54 employees include the 35 vehicle users, 9 truck drivers for whom truck assignments are unknown & rest being users for whom exact movement data is unknown. Finally geospatial shape files of Abila City and Kronos Island along with a 1:1 scaled map of Abila city along with prominent location marking is also provided.

In computing spatial temporal associations between employees, we first address some uncertainties we find in the data. The GPS logs provided for certain employees indicate these employees are active at given times, however their exact locations w.r.t. other employees are imprecise for computing associations. Further, though we are certain on locations of employees in transaction logs, length of time spent by the employee at that location is unknown. To address these concerns we build an Event-Information framework, as shown in Figure 1.

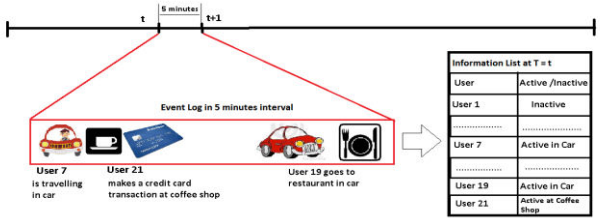

Fig. 1 Event-Information Framework for Synchronizing Different Data Types

Using this framework we compute the association between employees over user defined decision intervals. Within these decision intervals we first record all occurring events irrespective of their types into an event log & finally at the end of the interval update an information list indicating last known locations of all employees, assuming synchronicity between both GPS & transaction clocks. In transaction logs we assume length of time spent by the employee at the location to be equal to the mean time spent by a car user visiting the location. The mean time spent by a car user at each location is obtained by mining the GPS logs for visits at the location across car users, an e.g. of this is shown in Figure 2a. For synchronizing data we create dummy events of the transaction event over the assumed time spent.

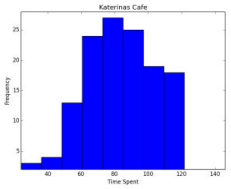

Fig 2a Distribution of Time spent by Car Users at Katrina's Kafe

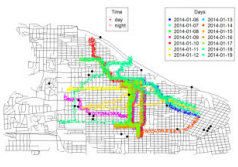

Fig 2b Jittery GPS coordinates reported by User 28 across all days

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

III. METHODOLOGY

The number of information lists we generate post synchronizing 2 weeks of data is a linear function of the length of the decision interval. As aforementioned this decision interval is user configurable and in setting this, authorities need to tradeoff between information loss and association time significance. These information lists indexed over time can be mined for different outputs to find suspect elements.

For finding employee spatial temporal clusters, we mine the information lists to first build an association matrix. Each element in the association matrix represents the strength of association of employee 'i' w.r.t. employee 'j', weighted differently over the event types to account for data imbalance. Associations between employees occur within decision intervals if the Greater Circle Distance (Haversine Formula, [3]) between the employees is less than a preset distance value. Also we weigh associations occurring during night hours separately. The association matrix as a result is not symmetric but a directed graph that is used to find employee clusters by applying a graph clustering algorithm on the association matrix; we use a version of the Markov Cluster Algorithm [4, 5] for this purpose.

Instantaneous departures from spatial homogeneity are indicative of suspect behaviors. We monitor the spatial homogeneity over the decision intervals to check for such aberrations. The information list available at the end of each decision interval provides locations of all employees and we use the Ripley's L-factor [1, 2] to measure spatial homogeneity. The L-factor would be close to zero, if employees are distributed in space and the value becomes negative when users are concentrated in an area.

On the visual layer we connect these outputs using common links and references. For e.g. in case of certain employees coming together in clusters we next look for spatial hotspots of where these associations are happening. We verify our observations and strengthen leads generated from these results by monitoring both L-factor and activity strength over time.

We also use the transaction logs separately to generate leads by creating a simple plot of employee credit spends to loyalty earnings over two weeks. This plot is also available at employee cluster level to monitor cluster spends.

IV. RESULTS

A dashboard as shown in Figure 3 has been developed to aid authorities in identifying both suspect elements and locations. The current version of the dashboard has been derived for a 5 minute decision interval over 2 weeks and preset value of 200 meters for determining associations. The MCL implementation is done outside the visual layer and only results are displayed. For accessing different outputs across screens common handlers between data files are also used on the visual layer. The entire development of the visual layer has been performed using Tableau.

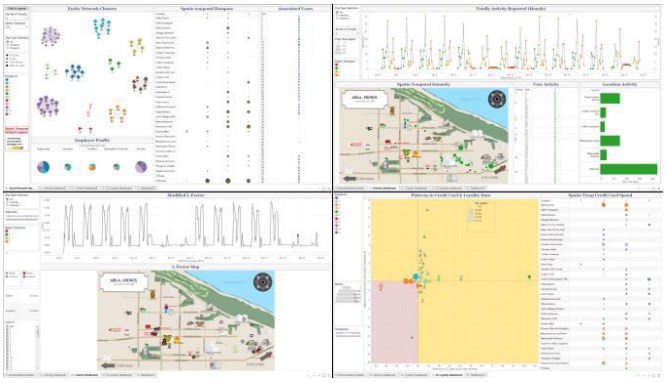

Fig. 3 Tableau Dashboards Using Data Mining Outputs

Mining through these screens we found employees 1, 10 & 23 to be clear suspects visiting Kronos Mart together at odd times and spending unusually high number of hours too. There are other users too visiting Kronos Mart independently at odd hours of the day.

V. SUMMARY

A combination of data mining techniques and visual analytics can help in understanding the data better and help identify regular patterns and abnormalities in data

ACKNOWLEDGMENT

The authors would like to thank Sreeja Arunkumar from Robert Bosch for her expert inputs on working with GIS files.

REFERENCES

[1] Ripley, B.D. (1976). "The second-order analysis of stationary point processes". Journal of Applied Probability 13: 255–266.

[2] Dixon, Philip M. (2002). "Ripley's K function". In El-Shaarawi, Abdel H.; Piegorsch, Walter W. Encyclopedia of Environmetrics. John Wiley & Sons. pp. 1796–1803. ISBN 0-471-89997-6. Retrieved April 25, 2014.

[3] Greater Circle Distances – Haversine Formula, http://en.wikipedia.org/wiki/Haversine_formula

[4] Van Dongen, S. A Cluster Algorithm for Graphs. (2000) Stichting Mathematisch Centrum, Amsterdam, Netherlands

[5] MCL Algorithm, http://www.micans.org/mcl/