Developing web-based data analysis tool for investigative analysis using R and Shiny

Kelly Koh Kia Woon
Singapore Management University
kelly.koh.2020@mitb.smu.edu.sg

Manmit S/o Narmal Singh
Singapore Management University
manmits.2020@mitb.smu.edu.sg

Syed Ahmad Zaki Bin Syed Sakaf Al-attas
Singapore Management University
ahmadzaki.2020@mitb.smu.edu.sg

ABSTRACT

xxxx

1. INTRODUCTION

Almost every action leaves a digital trail. Major technological shifts in the past decade have made the collection of digital evidence, such as GPS records and payment transactions, a significant tool in criminal and civil investigations[2]. It is crucial that law enforcement agencies could use and transform the ample data into insightful information to aid their investigations. The growth in volume and variety of digital data increases the time and resources needed to analyse them. Furthermore, cases increasingly the analysis of multiple devices followed by the correlation of the found evidence.

However, most law enforcement personnel lack training in data science. Without effective data analysis, investigators will struggle to find relevant information for their cases. And the results of the digital evidence analysis can be challenged in court when not well-represented. Given that most of the work in digital forensics involves association to individuals, law enforcement agencies need forensic tools to approach identity management and representing a person's network. Work in this area will allow for the identification of likely suspects and other anomalies. By integrating interactive visualization with automated analysis techniques, digital data can be presented in meaningful ways and allow investigators to interactively guide the investigation.

Addressing this challenge, we designed and developed , a __ dashboard to allow law enforcement agents to conduct investigative analysis of GPS and credit card data via an open-source tool. It aims to simplify data exploration and analysis to gain valuable insights into the behaviour of individuals quickly. The study is built upon VAST Mini Challenge 2's task to find unusual patterns in GASTech employees credit card records and GPS tracking records of their cars. We demonstrate the potential of ___ with interactive maps to explore GPS paths, interactive network graph to explore relationships of _____, UpSet plot to understand co-location of individuals and one-way ANOVA to determine differences in spending behaviour. This paper documents our approach to design and develop the interactive application targeted at law enforcement agencies. This introduction is followed in Section 2 by an explanation of our objectives and motivation. Section 3 provides a review of existing techniques used to visualize GPS and credit card data. Section 4 details the data used, and design framework used. Section 5 provides a visual overview of the application. Section 6 summaries the findings from the use case. Section 7 concludes the report and offers ideas for further development.

2. MOTIVATION AND OBJECTIVES

Our research and development efforts were motivated by the general lack of effective and easy to use web-enabled client-based visual analytics tool for investigators to discover patterns in digital data such as GPS and credit card transactions. The interactive tool aims to address the following analysis requirements: - Are the locations and individuals statistically different in mean spend? - How many visitors went to location A, B, C and D? - What is the visit frequency of a specific staff to a particular location? [heatmap] X - Can either the staff or the locations be clustered into groups, based on similarity? [heatmap] X - How long did a staff stay in a particular location on a particular day? [map] X - Were there any staff who exhibited similar movement patterns? [map] X - [visnetwork] Which location connects the individuals

3. REVIEW OF EXISTING TECHNIQUES

3.1 Visualizing Set Data

Understanding relationships between sets is a common visual challenge that many analysts face. Venn diagrams has been the traditional choice of graphics adopted to make sense of sets data. However, as the number of sets increase, the Venn diagrams becomes too complex to and puts burden on the readers. An investigator who would like to see the size of the co-location of individuals of more than 4 locations will struggle without a proper tool. For example, R package such as 'Venn' can only draw and display up to 7 sets, as shown in Fig 1, and the results take time to interpret.

Figure 1: A Venn Diagram showing the use of 'Venn' R package to show the intersections of 7 sets To overcome this problem, we can adopt the visualisation technique using 'UpSetR' package. This will help us breakdown the information into parts: i) visualizing the total set size and ii) visualizing the intersect size.

3.2 Network Graph

Leveraging visuals such as scatterplots, histograms and bubble charts, relationships between entities in datasets can be fleshed out. These charts allow for a quick overview and identification of relationship patterns. In addition, statistical methods such as correlation, can denote the strength of relationships between quantitative variables. Specifically, using Cohen's d to measure the difference in standard deviation between two groups in a population, and Pearson's r to measure the strength of a linear association (Price, 2012). As an example, the scatterplot below shows the moderately negative relationship between temperature against elevation using Pearson's r .

Figure

While these methods are useful, they are not sufficient to visualise complex relationships between elements. To thoroughly visualise and investigate relationships, Network Graphs should be used. Network Graphs are powered by graph theory, that transforms the entities and connections within datasets (Carlson, 2020) to nodes and edges respectively. These graphs can flesh out i) relationships between all the entities within the dataset; ii) number of connections each entity has; and iii) importance of relationships that connect the entities. In addition to visualisation, Network Graphs can be statistically analysed to highlight underlying relationships using methods such as, i) degree centrality – the number of ties to a node, and ii) eigenvector centrality – the node's importance within the network (rdrr.io, 2020). Network Graphs are often used to analyse transportation and social media networks.

4. DATA

Our data was obtained from Mini-Challenge 2 within the annual IEEE Visual Analytics Science and Technology (VAST) Challenge 2021. It has provided vehicle tracking data and credit card purchases of the GASTech employees 2 weeks before the disappearance of several employees to support the investigation in csv format. Data was cleaned and transformed using in R using base R and the dplyr package. These activities are performed by server side of the application and hidden from the user.

5. THE APPLICATION

5.1 System Architecture

___ is an interactive tool developed using Shiny, an R package used to build interactive web apps and various R packages. We used Shiny to deploy the plots as its open-source nature ensures a modularized and reproducible workflow that can distributed easily. The reactive programming model in Shiny allows users to interact with customisable widgets and view the changes to the visualizations from their input change. The design of the Shiny web app flowed from the key questions of the data: i) First tab – ii) Second tab – iii) Third tab – iv)
v)

5.2 Data Visualisation

5.2.1 ANOVA

One-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent groups. The ggbetweenstats function from ggstatsplot package was used visualise the output of ANOVA tests. The function supports the most common types of hypothesis tests, including Welch's one-way ANOVA for parametric data, Kruskal–Wallis one-way ANOVA for non-parametric, Fisher's ANOVA for Bayes Factor. It also overcomes the omnibus test nature of ANOVA by selecting the appropriate post-hoc pairwise comparison tests and showing results directly within the plots.

The app enables the user to select the groups that they would like to explore and the univariate distributions for each group are displayed in the panel based on the bin size selected. Based on the understanding of the data distribution, users can then select the appropriate test statistics (including parametric, non-parametric, robust and Bayes Factor) for the ANOVA test.

5.2.2 UpSet Plot

UpSet plots overcomes the set size limitation of Venn Diagrams with a novel way to view set data by the size of their intersections. It avoids the visual complexity of Venn Diagrams and focuses on communicating the size and properties of the set aggregates and interactions with bar charts. The UpSetR package was used to build the Upset plot in the app. This package requires the dataset to be in binary matrix format, with the columns representing the universe of sets. It allows users to change visualization attributes to explore the intersections between the sets.

To facilitate data exploration, the app provides users with the option to select the sets of interest, select the method of ordering of intersections and showing empty intersections. User can make use of the chart to explore the size of each set combinations. Format settings are also included to allow users to modify the aesthetics so that it can be readily included in reports.

5.2.3 Heatmap Dendrogram

A heatmap would allow for easy visualisation of frequencies between two categorical variables, while a dendrogram would allow for better analysis of similarities within these same variables. To achieve these aims, the Heatmaply package was employed to marry both the heatmap and dendrograms into a single interactive visualisation. The app allows users to customise the dendrogram settings, such as its seriation, clustering, and distance method, while at the same time, enabling the user to interact with the heatmap via the use of its zoom and tooltip functions.

5.2.4 Spatial Map

Maps are usually the go-to visualisation for spatial data. It allows the user to interact with the map to understand both movement trails and proximity of locations from each other. For this, both the sf and tmap package was employed.

The sf package provides users with a comprehensive list of spatial data manipulation. Separately, the tmap package allow users to visualise these spatial data on a map, along with its related map functionalities.

5.2.5 *Network Graph*

Unstructured and structured data can be used for Network Graphs, and the datasets must be carefully prepared into nodes; and edges data. Minimally, the nodes dataset contains columns which i) "id" - a unique identifier for the node to the edge and ii) "label" - the name of the variable. The edges dataset must contain at least i) "id" – maps the node to the edge data; ii) "from" (source); and iii) "to" (target) columns. The "from" and "to" columns map the relationship between the source and target. In R, the visNetwork package provides the dexterity to customize the network graph, such as, its layout, node image, edge colour and opacity, font size and edge direction (Thieurmel, 2019). Credit card and loyalty card data was used in the package to explore relationships between i) credit card expenditure-location, and ii) loyalty card expenditure – location. The app has features for the user to customize the network display and understand the networks' eigen and degree centrality. The default layout is force-directed, which tries to produce a nice visual. The graph has a zoom function and nodes can be dragged to explore specific points-of-interest ("POI").

6. CASE STUDY: VAST MINI CHALLENGE 2

The final interactive plots in the published web application, and an illustration of the potentially wide range of insights that can be gleaned from the application are briefly described below.

We demonstrate the testing of difference in mean spend across locations and credit card users using the ANOVA plot.

Fig _. Interactive ANOVA User can select the locations of interest from the main panel after exploring the data in the 'EDA' tab. 'Brew've Been Served,' 'Jack's Magical Beans,' Kalami Kafenion' and 'Guy's Gyros' are selected by default. If the univariate distributions from the histograms reveal skewedness, the user can make judgement to select a non-parametric test. The visualization automatically picks a Kruskal-Wallis test at 95% confidence level. The ANOVA results shows that there are significant differences between the spend across the locations ($p < 0.01$). The Dunn test is the automatically selected as the post-hoc pairwise test, which displays significant comparisons on the chart. Of which, results show that 'Brew've Been Served' and 'Guy's Gyros' have significantly different spend ($p < 0.01$). This provides evidence for the users to support the claim that these locations are of different nature.

Next, we demonstrate the exploration of co-location of individuals via UpSet plot.

Fig _. Interactive Upset Plot The UpSet plot allows user to – at a glance – gather information about the number of visitors to a location and how the size varies when intersecting with other locations. For example, if a user would like to know how many of the visitors who visited the Golf Course also visited the other retailers, one can select the locations in the side panel and order the intersections by degree. The plot shows that of the 5 individuals who visited 'Desafio Golf Course,' 1 of them also visited 'Roberts and Sons' and 'Albert's Fine Clothing' and 2 other visited Albert's Fine Clothing and 'Shopper's Delight.'

Fig _. Interactive Network Graph Thicker lines and nodes clustered close to each other indicate popular locations and closer relationships. Selecting a node highlights immediate relationships, from which the user can further explore secondary relationships and connection paths. Users can seek out specific networks by searching credit card, loyalty card and location POI. Selecting a POI fleshes out related nodes within the network. This would provide a sense of informal relationships within a network by location. In addition, inferential statistics analysis is done when the user selects eigen and degree analysis. This highlights the important and popular relationships respectively within the POI's network. Important relationships include nodes that have few but well-connected connections. The node colours fade in tandem with relationship importance. Popular relationships are nodes with many connections.

7. FUTURE WORK

The prototype has immense potential for interactive and analytical enhancements. The application could provide users an interface to upload, clean and step-by-step guide to transform data files into appropriate format for each visualization.

Furthermore, the application could be enhanced to include brushing-and-linking of different types of charts and embed drilldowns to provide details on demand.

The tmap package can be enhanced to link up to actual images of road and environment (e.g. Google Street Map) to allow investigators to explore the surroundings of locations.

The visNetwork could be improved with additional options for analysis. This includes centrality analysis options such as betweenness and community detection.

References

[1] Garfinkel, S.L. 2010. Digital Forensics Research - The Next 10 Years. 7, (2010), S64–S73.

[2] Goodison, S.E. et al. 2015. Digital Evidence and the U.S. criminal Justice System - Identifying Technology and Other Needs to More Effectively Acquire and Utilize Digital Evidence. 890, (2015).