Case Study of Dino Fun World Movement and Communication Data

Jordan Riley Benson*, Rajiv Ramarajan, Nascif Abousalh-Neto, Paul Vezzetti

SAS Institute

ABSTRACT

We present our solution to the VAST 2015 Mini-challenge 1. In our solution we utilized existing visual analytics software and custom tools to analyze the movement and communication data provided for the fictitious Dino Fun World amusement park. Our process focused on collaborative data discovery and the application of analytics procedures. In this paper we outline the techniques we used and highlight both the successes and shortcomings we found in our toolset when applying it to this challenge.

Keywords: Information visualization, visual analytics, big data, data visualization, exploratory data analysis, VAST.

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation-User Interfaces; H.1.2 [User/Machine Systems]: Visual Analytics;

1 INTRODUCTION

The problem of analyzing large amounts of movement and communication data is common to many fields. It is easy to imagine the techniques used for this challenge being extended to urban surveillance or manufacturing and distribution optimization.

2 PROCESS AND TOOLS

Our process involved moving the data back and forth between custom Python scripts and SAS Visual Analytics[1]. We used the collaboration tool Slack[2] to communicate within our small team and share findings as well as distribute additional data.

2.1 Data Preparation

As the data size was not unreasonable for manipulation on a typical computer we did the majority of the data preparation using Python scripts. The data sets for the three days were assembled, joined, and cleaned before being used.

Some of the data was captured in images provided on the park website including the park layout, the names of the attractions, and the regions of the park. Those were either extracted by tracing the image with a vector editor or through manual transcription.

Much of the provided data needed to be enriched with additional metrics to support more in-depth analysis. For example, check-ins had to be associated with the attraction based on location alone and other metrics like the time spent at the attraction had to be calculated.

Of note, we automated the process of mapping check-in locations to attractions by mapping the check-in coordinates to the polygonal outline of that attraction by finding the shortest distance to any edge of the polygon, which was assumed to be the likely location of an entryway.

* Riley.Benson@sas.com

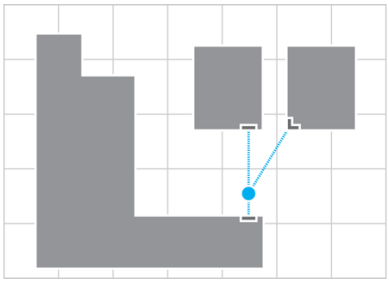

Figure 1: Method for mapping position to likeliest attraction.

2.2 Visual Analysis

Once the data and additional metrics were prepared we uploaded it to an in-memory distributed server to make rapid visualization easier. SAS Visual Analytics was used to examine the distributions of the metrics, find visitors that were outliers, find temporal patterns, and identify high frequency paths.

The output from this investigation was a set of time ranges, visitor ids, and attractions of interest that were then examined in more detail to assemble stories.

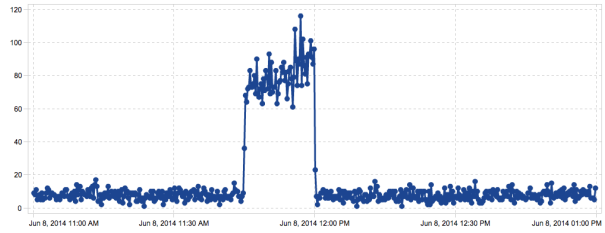

Figure 2: Time series of communication frequency.

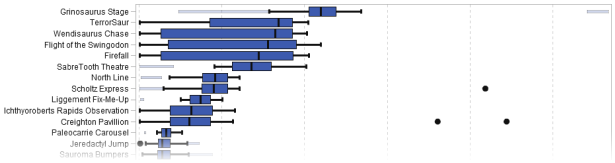

Figure 3: Box plots of check-in durations per attraction.

The path analysis feature in SAS Visual Analytics was primarily intended to show how customers move through a company's process. Call center flows or e-commerce website traffic for example use this type of analysis often. We repurposed it to find groups of visitors that traveled together throughout the day.

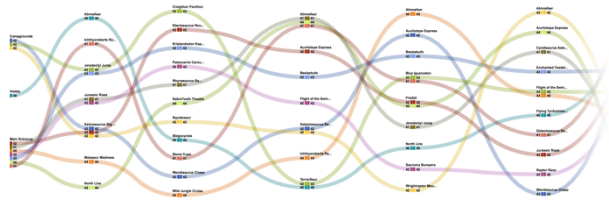

Figure 4: Path analysis output by attraction for large groups.

The output from the path analysis is visualized as a Sankey diagram, but this method does not scale to the many thousands of visitor paths that exist in the data. To reach interpretable results we had to only display paths that met specific criteria and use the groupings of visitor ids to drive further investigation.

2.3 Additional Visualization

Our existing tools did not have support for rendering on top of a provided geospatial map background so we created these with custom scripts. Scatterplots and animated heat maps were straightforward to create but key to putting anomalies in their spatial context.

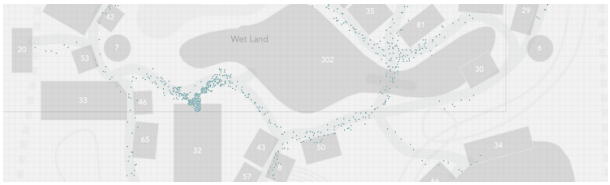

Figure 5: Scatter plot showing visitor location during anomaly.

Rendering a single individual's path throughout the park was an unexpected challenge. Visitors tended to walk across the same path many times and that generated a lot of over-plotting. To help counter this we introduced jitter to each of the movement points for a visitor within a five-meter square to reduce the chance of overlap and to also include some representation of the positional uncertainty present in the data.

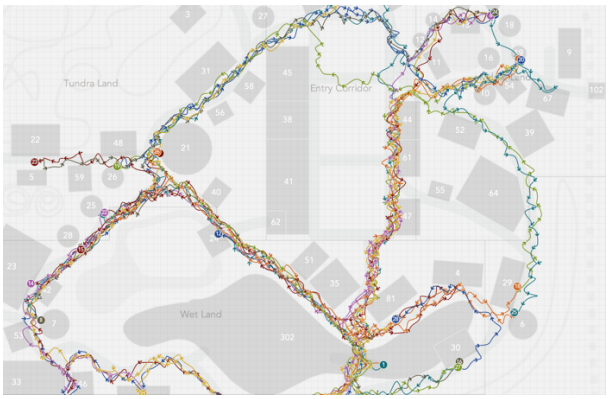

Figure 6: Visitor path through park shown using directed linesets.

The jittered points were connected using Bezier curves that remained continuous throughout the reported movement points in the path being visualized. The intent was not to draw the path as the visitor might have taken it, but instead to provide a curve that was easy for the eye to follow along a leg of the journey. This relied on the same concept used in the LineSets[3] visualization. Arrows were added to show directionality and segments were divided based on check-ins and color coded for easier identification.

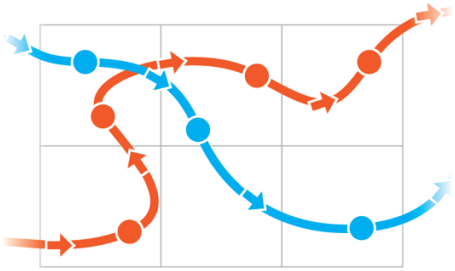

Figure 7: Diagram of the visitor travel paths visualization.

3 Interpreting Results

Combining the high level results from our analytics with the detail visualizations we were able to identify visitors of interest related to the park vandalism but were unable to uncover anything other than circumstantial evidence showing that they spent a great deal of time at the Creighton Pavilion sometime Sunday morning. The crime was first discovered by the general public at 11:45 AM and the park authorities seem to become aware of the crime at 12:00 PM. We based these assumptions on large communications spikes in the data involving visitors clustered near the entrance to the pavilion.

We focused on describing broad types of visitors that had similar attraction attendance in the park instead of focusing on groups that moved in similar ways. This approach allowed us to describe possible improvement to the park in terms of visitor preferences. We assigned the names The Masses, Audience Members, Thrill Seekers, Wanderers, and Small Fries to the clusters that were found by examining attraction type frequency distributions. We also identified visitors that were members of a very large group that traveled together and referred to them as Herds.

3.1 Potential Improvements

We were unable to fully answer several questions due to the absence of key analytics or visualizations that we did not have time to implement.

The largest shortcoming was our inability to perform proper geospatial clustering. The path analysis method we used was able to detect groups that checked in at exactly the same attractions through the entire day but is very inflexible. Group members that decided to visit a single attraction separately before rejoining than the rest would be excluded for example. Also, without true clustering on the movement we were unable to detect common movement patterns throughout the park that didn't rely on the order of attraction check-in. That form of clustering would have allowed us to aggregate and display the the primary paths at various times through the weekend.

Another weakness was the lack of custom mapping support in SAS Visual Analytics. This required us to extract the data sets generated by the analytics procedures so we could plot them using our custom visualization scripts. This slowed down the iterative data discovery process. Being able to provide an image backdrop behind of the axis based graphs would be useful for other problems as well.

Acknowledgements

Many thanks to Matthew Horn for applying his voice talents to our video presentation:
https://www.youtube.com/watch?v=BIPscav27Js

This work relied on assistance from Steve Clark who provided IT support for the software we used for this challenge.

References

[1] SAS Institute. SAS Visual Analytics. [Online]. Available: http://www.sas.com/en_us/software/business-intelligence/visual-analytics.html

[2] Slack. [Online]. Available: https://slack.com/

[3] Alper, Basak, et al. Design study of linesets, a novel set visualization technique. *Visualization and Computer Graphics, IEEE Transactions* volume 7.12, pages 2259-2267, 2011.