Applying Visual Analytics to Explore and Analyze Movement Data

Eren Cakmak Alexander Gärtner Thomas Hepp Juri Buchmüller Fabian Fischer

Daniel A. Keim

Data Analysis and Visualization Group*
University of Konstanz, Germany

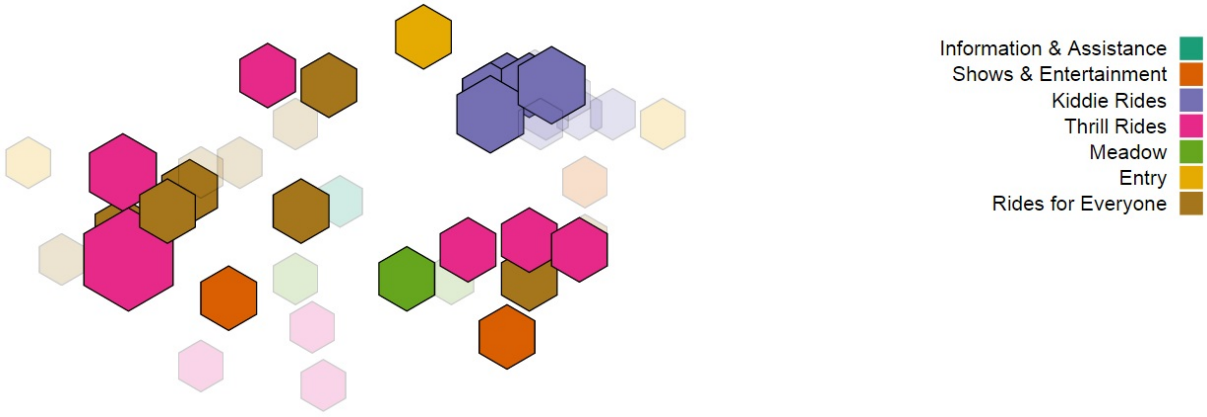

Figure 1: Overview of attraction visits for one visitor of a single day. Each hexagon represents one attraction. The color describes the attraction category. Only the highlighted attractions were visited.

Abstract

The VAST Challenge 2015 movement dataset is mirroring current challenges in the analysis of large spatiotemporal datasets. We present a tool featuring different exploratory approaches analyze and visualize spatiotemporal data to build and confirm hypotheses. Our tool helps the user to find patterns, anomalies and groups in a data set that can not be processed manually. We present custom visualizations to solve the tasks stated by the VAST 2015 Mini-Challenge (MC1).

Index Terms: H.2.8 [Database Management]: Database Applications—Data mining; H.5.2 [Information Interfaces]: User Interfaces—Graphical user interfaces (GUI)

1 Introduction

This paper completes our submission for the VAST 2015 Mini-Challenge 1 (MC1). The challenge scenario "Mayhem at DinoFun World" features movement and communication data of the visitors of a fictitious amusement park over the course of one weekend. A football star, called "Scott Jones", was visiting the park. He had several events over the weekend showing his football moves. As described in a news article, there was an incident where people rioted at the Creighton Pavilion attraction.

Over the weekend, about 11 000 people visited DinoFun World. For this large amount of movement data, it is difficult to find groups moving around together or to define groups that share a similar behavior or show suspicious patterns. The following paper shows approaches how we handle the movement data in order to derive semantic knowledge using a visual analytics approach.

*e-mail: firstname.lastname@uni-konstanz.de

2 Preprocessing

To effectively analyze the movement data, preprocessing has to be applied to deal with corrupt or invalid entries. Using the data mining tool KNIME[1], we found out that the data for Friday is corrupted, as after 8:12pm, no more data is available, while communication data is available until 11:25:54pm. Moreover, on Saturday and Sunday, data for the respective time span is available. The rest of the data is consistent with the exception of one missing value on Sunday, that we ignore. Altogether, there are 25 361 394 valid data records, which is too much data to effectively visualize every single track, because of hardware limitations, limited display space, and the resulting clutter. Therefore, clustering is required to reduce the complexity of the dataset.

3 Clustering

In order to find groups of similar behavior, the data needs to be clustered. Using Tableau[2], we found out that one clustering approach is to identify various features describing a group moving as discussed in Section 3.2. Besides, Symbolic Aggregate Approximation (SAX)[3] can be applied as an alternative clustering approach.

3.1 SAX-Based Clustering

Our approach uses clustering on different time series: Looking at the movement of a single person, we have a time series in a two dimensional space. We mapped the 100x100 pixel map of the park to a one-dimensional space. This way, we represent the movement

[1] http://www.knime.org/
[2] http://www.tableau.com/
[3] https://code.google.com/p/jmotif/wiki/SAX

as a time-series. Choosing the right time interval is crucial, as a value too high would mean a lot of lost movement data, whereas a low interval results in longer computational time. For these reasons, we preferred feature based clustering over a SAX approach in the end.

3.2 Feature-Based Clustering

Using KNIME, we calculated features for every visitor, identified by unique IDs, visiting the park: (i) total number of movements per day, (ii) total number of check-ins per day, (iii) time spent in park, and (iv) check-ins for attractions grouped by categories (e.g., kiddie rides).

We tried to identify groups of people using EM-Clustering, DBSCAN and k-means clustering approaches. DBSCAN yielded the best results. Then we visualized the result of the PCA shown in Fig. 2.

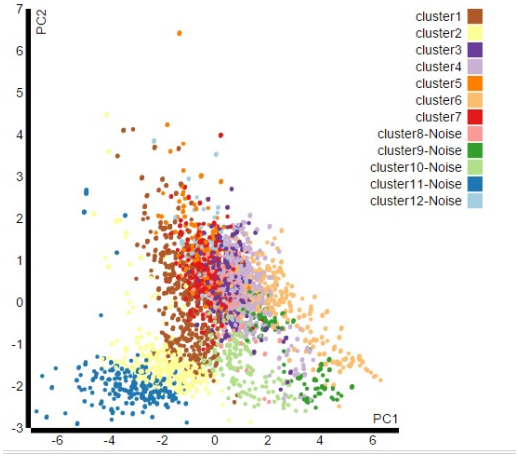

Figure 2: Principle Component Analysis (PCA) result for Sunday. Each point represents a visitor and the color is mapped to the closest cluster described in Section 3.2.

4 Visualization

As mentioned before, data for movement and communication of visitors was available. To identify visitor groups and anomalies over the available time span, we applied four visualizations.

Hexagons – Fig. 1 visualizes the attraction visits for each visitor. By this approach, general patterns in visits are found, e.g. the so-called "Thrill Ride Junkies". In our tool, visitors are selected by either a list sorted by the calculated cluster or in Fig. 2 by the mouse-over interaction (e.g., selecting an outlier). The visits to each attraction over the day are shown in a star glyph via a mouse-over effect on the hexagons.

Heatmap – Using heatmaps, we analyzed the attraction visits and the movement in the park to find unusual patterns. We were able to determine the time of both shows of the VIP with the heatmap for the attractions and we discovered the missing second show on Sunday. This finding and the closing of Creighton Pavilion on Sunday after 12 noon led us to the conclusion that the crime happened on Sunday. The movement heatmap shows the number of movements for a specific location sensor aggregated in a ten minute interval. Using the heatmap and the communication data visualized in a multi line chart, we were able to find events and hot-spots as shown in Fig. 3.

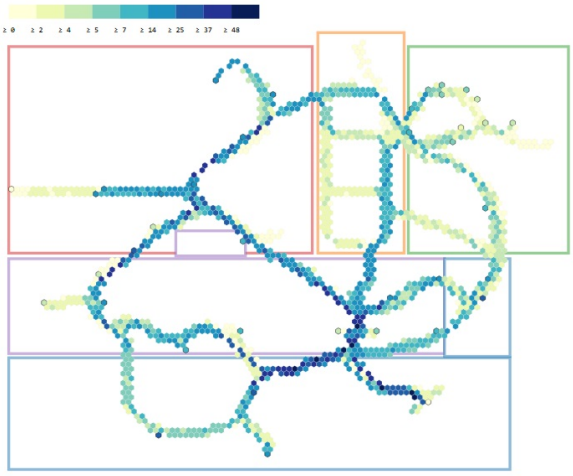

Figure 3: Heatmap showing movement data on Sunday 11:00 pm. Each Hexagon represents a sensor and the aggregated movement amount for all visitors in range of that sensor.

Chord diagram – Using the Chord diagram, single visitors can be selected and highlighted to show their paths through the park, finding patterns or anomalies in the movement. Each segment of the circle border represents one attraction.

5 Findings

Below, we present selected findings. The complete list of findings is part of our submission for the VAST 2015 Mini-Challenge 1.

Groups – We identified 11 groups or clusters over the weekend. These vary in number of group members, visited attractions, time spent in the park or movement speed. One of the groups is called the "Thrill Ride Junkies", consisting of 3.7 visitors on average and they move around a lot in the park mainly visiting the thrill ride attractions.

Patterns – One pattern is at "Liggement Fix-Me-Up", which is normally visited at a very low rate expect the spike on Friday at 1-3 pm. Another pattern is the noticeable decrease of check-ins at the "Galactosaurus-Rage" ride on Friday between 7:00 and 8:00 pm as the attraction is normally visited with a high frequency.

Anomalies – A crime related anomaly is the closure of the "Creighton Pavilion" on Sunday afternoon while it normally is open until the evening and the missing show at the "Grinosaurus Stage". We found a suspicious visitor who is entering the park at Friday and disappearing at the "Ichthyoroberts-Rapid" attraction, reappearing on Saturday at the "Scholtz-Express", from where he walked straight out of the park.

6 Conclusion and Future Work

With our visualization we found various groups and anomalies in behavior of visitors and general patterns. To improve the explorative nature of the tool, more interaction and filter options should be available to support the user. To use the grouping with hexagons more efficiently and to exploit the advantages of visual appeal and representational accuracy mentioned by Carr et al. [1] of a hexagon geometry, a layout algorithm is required which abstracts the map.

References

[1] D. B. Carr, A. R. Olsen, and D. White. Hexagon mosaic maps for display of univariate and bivariate geographical data. *Cartography and Geographic Information Systems*, 19(4):228–236, 1992.