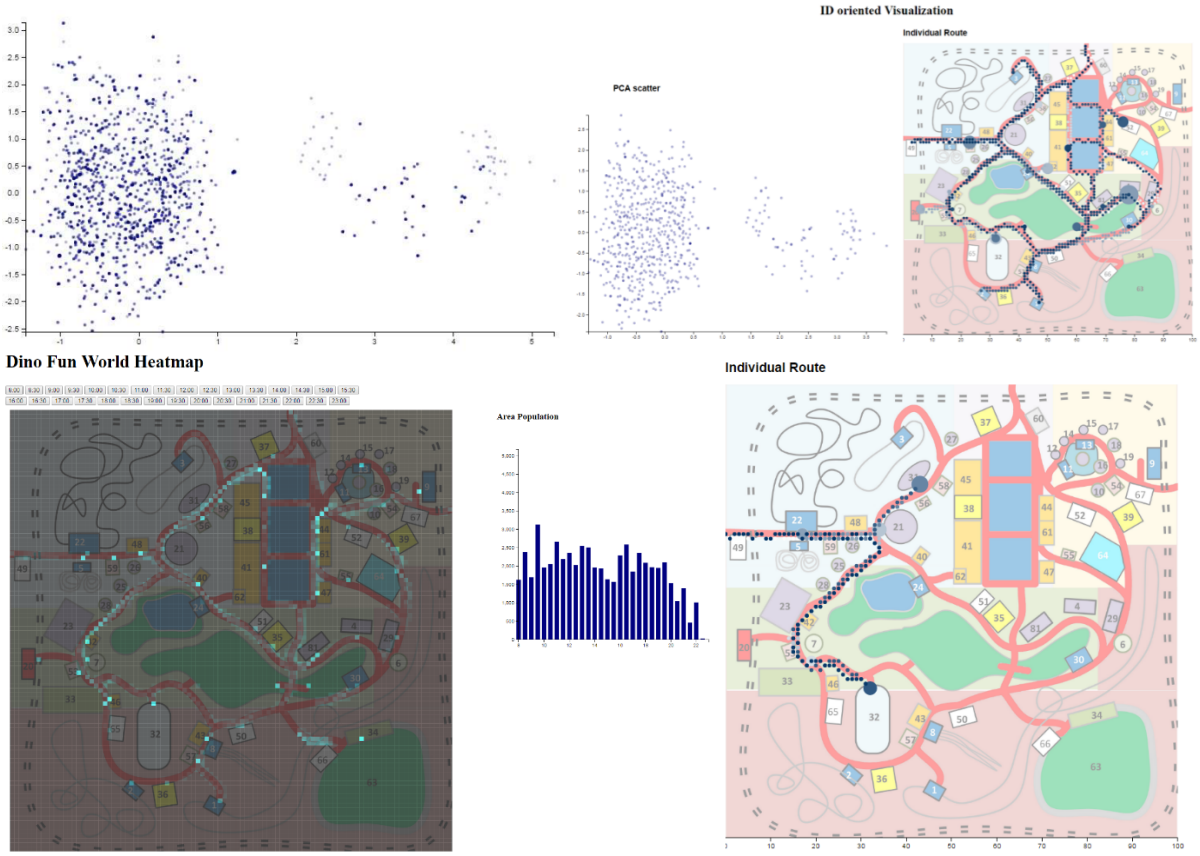
Visitor Oriented and Place Oriented Visualization of Movement in an Amusement Park

Xiangci Li, Zihuan Zhang
New York University Shanghai



From left to right, top to bottom, Figure1: PCA scatter plot of all data. Figure 2: Visitor Oriented Visualization with reduced PCA plot on the left and individual route on the right. Figure3: Region Oriented Visualization with heat map on the left and region population histogram on the right. Figure4: The route map of a potential suspect.

Abstract

The record of visitors' movement across one day can produce massive data. We designed visitor oriented visualization to visualize the moving trajectory of each visitor and find out anomalies. We also designed place oriented visualization to visualize how the population distribution across park change throughout a day.

1. Introduction

This problem is taken from VAST 2015 mini challenge 1. DinoFun World is a simulated amusement park with size 500m * 500m. All visitors are equipped with a park app to check in/out the park or go on a ride. There are sensors all around the park which is sensitive within a 5m*5m area. On a weekend of June 2014, a crime happened in DinoFun World. A pavilion which exhibited the memorabilia of Scott Jones, a soccer celebrity, was destroyed. The park wants to find out the potential suspects. The resources the park has are the real-time location (every 1 min) of all the tourists of this weekend, each with a unique ID number; a map which gives all the

information of rides, restaurants and pavilions; and a website contains basic information of DinoFun World. However, there are millions of data recorded in the form of a table which makes it time-consuming to track every individual. If we can cluster these tourists into certain category and try to find some abnormal pattern, it will greatly narrow the scale of potential suspects. As visualization gives human the direct intuition of clustering, we think visualization can be a great tool to improve efficiency. Furthermore, visualizing tourists' movement can also help the park to find out which rides the tourists like the most, when a ride has the most of tourists, where exactly in the park are the most crowded etc. Therefore, our visualization can also help the park to better manage their human resources. All these reasons motivate us to work on this challenge.

2. Dataset

There are three csv files, storing the visitors' movement information on Friday, Saturday and Sunday. All visitors to the park (except for very young children) use a park app to check in to the park and rides and to communicate with fellow visitors. Specifically, each file contains around 6 million of tuples, which has time, id, movement type and position attributes. Each visitor has a unique id number. Movement type has two kinds of types: "check-in" and "movement". Position is encoded by x-y coordinates of grid cells. The park is equipped with sensor beacons that record movements within the park. Sensors are sensitive within a 5m x 5m grid cell. All pathways in the park are covered by these sensors, as are the ride check in locations. Locations are not recorded while people are on rides or inside attractions (including restaurants, stores, and rest rooms). The data is generally clean: no visible missing values, no apparent dirty values, and the format is unified. Structurally, the datasets are not complex, but the datasets are large and there are many patterns to reveal.

In addition, the park web site provides much contextual data. A map of the park illustrates the park layout as seen from an overhead perspective. This gives the footprint of rides and paths through the park. A map index and text descriptions of the rides and attractions are provided to give more information about the environment within which the park visitors are traveling.

3. Related Work

3.1 Data Processing Method

To find the suspect is a part of anomaly detection which has been widely studied and used due to its powerful application on monitoring activities in security systems. Chandola et al. define that anomalies are patterns in data that do not conform to a well-defined notion of normal behavior [8]. They have made an extensive survey on anomaly detection. One category of anomaly detection applies supervised learning technique to classify normal data and abnormal data [29]. However, labels are expensive to obtain. In the case of DinoFunWorld, even manually classifying abnormal movement is hard, therefore this method is not applicable here. To avoid the expensiveness of labels, anomaly detection with unsupervised learning method is proposed [9]. It finds out anomalies by assuming most of data are normal data or even does not require training data. Kohonen proposed self-organizing map, which is a is an unsupervised artificial neural network based architecture for anomaly detection [18]. The self-organizing map has the property of effectively creating spatially organized internal representations of various features of input signals and their abstractions. Munoz et al. proposed using self-organizing map on outlier detection [24]. K-means clustering algorithm is another powerful unsupervised learning method by doing clustering on data. The idea of k-means is first proposed by Steinhaus [28] and first named by MacQueen [23]. The standard algorithm is proposed by Lloyd [22] and Forgy [10] separately. However, the algorithm has NP-hard complexity, so many efficient implementations are proposed later [12, 16]. Because of the simplicity and good effectiveness, we will adopt k-means clustering algorithm as our main data processing algorithm. Furthermore, many studies are conducted on dealing with complex multi-dimensional data [1]. One of the effective way is using principal component analysis proposed by Hotelling [13]. Richardson provided a detailed explanation and the algorithm

of PCA [27]. Since the data provided by DinoFunWorld are multi-dimensional, PCA is particularly useful for data reduction before feeding data into k-means clustering algorithm. Moreover, Breunig et al. proposed using Local Outlier Factor, which is a degree that measures how isolated an object is with respect to the surrounding neighborhood, to detect outliers [3]. Thanks to the open source codes, the python codes of PCA [35] and LOF [36] are available online.

3.2 Visualization Designs

3.2.1 Movement visualization

One key component of this paper is the map of Dinofun World. Therefore, movement visualization is an effective way to show the pattern of the visitors.

Previous paper provides us with rich examples: Buchin [4] and Phan [26] worked on theories as well as the design of flow map. This is the most intuitive way of showing movements, especially in migration. Karnick [17] went one step further, he gave the detailed algorithm of route visualization and evaluation of it. These theories can be the solid basis of our visualization design.

3.2.2 Multidimensional data visualization

To deal with the multidimensionality of the given data, Inselberg [14] suggested parallel coordinates. This is a direct way to lay out multidimensional data on 2-D plane. However, for a large dataset the parallel coordinates will look crowded. M.Wish [19] solved this problem by multidimensional scaling which gives a statistic result on the similarity of data. These methods offer detailed principles on how to visualize high dimensional data. They also introduce some tools to help deal with multidimensional data more efficiently.

3.2.3 Design Principles

How to make sure a visualization is effective? Munzner [25] collected and organized a set of design principles as well as validations of them. These will be the theory this paper follows. Kandogan [15], based on these principles, designed star coordinates to show multidimensional data. The visualization combines effectiveness as well as the beauty. Some research also focused on human behavior. Cao et al. [7] designed TargetVue to detect anomalous users in online. The attributes Cao et al. chose to define a user is very useful in define a visitor in this paper. TargetVue demonstrate effectives ways for people to detect outliers which are hard for computer to detect. Buja [5] focused on another aspect on visualization design: interaction. This paper shows some methods on interactive visualization which enables user to explore the data and focuses on their own interests. In this paper, interaction should make the visualization more centered as there are lots of data which may be noises when user want to focus on one certain dataset.

3.2.4 Previous paper on this problem

To come up with an appropriate design, we refer to the previous paper to get inspiration.

Wei, Shuang [32] processed the raw data creatively. Instead of organizing the data by ID they try to identify group who travels together. This is an important discovery because individual behavior is affected by group. For example, a group with a child will wait outside the children zone, even they cannot play inside.

Steptoe [30] provided 2 clustering methods and 4 visualization design. The heat map approach and calendar approach are very special ways to solve this problem. The network analysis also demonstrates new information hidden in the raw data.

Some paper also uses trajectory visualization [11]. This can be the base of our visualization but when there is a large dataset, this tool become unable to identify. This is a potential aspect for improvement.

Wang [31] suggested spectrum as a way to cluster visitors. The color pattern it demonstrates is very effective. If the spectrum can be combined with trajectory visualization, an intuitive and clear visualization can be designed.

Benjamin [2] took time into consideration. He clustered visitors into groups based on location and time. These attributes can be added when we analyze visitor's movement pattern.

Some people creates a system which contains several visualization tool at a time [33]. Yu's paper is a useful database to choose a visualization which people perceive the most information from. Some of the visualization can

also be combined to have a better understanding of data.

There are also bad examples that may confuse user [6]. Although the hexogen and the color is fancy, they covey very little information. Therefore, it increases the paper ink but reduce the paper ink ratio. This is a very inefficient and confusing.

4. Design and Methodologies

Since there is huge amount of high dimensional data, for data preprocessing, we first used Python scripts to rearrange the raw data into each ID's route of movement along time. We used Numpy library [37] and Jupyter Notebook [38] to help process massive numerical data in the form of matrices. We also map each facility's and big area's position to the coordinates in the 100*100 map to regenerate the missing map data. Based on the objective of our task, the whole design is divided into mutually supportive two schemes: people oriented and place oriented approach.

4.1 Visitor oriented approach

We made a Python class "Person" to process the raw data. We crawled the row data line by line and stored them into a Python dictionary of Person objects. Each object stores a visitor's movement history. Then we calculated the features of each id, including the absolute time visited and the time spent in each area and each facility, number of total check in occurred and the time they entered and exited the park. Based on these extracted features, we used PCA to reduce the number of data dimensions into two. Then we applied k-means algorithm to cluster the two dimensional data. We plot each id into a scatter plot (Figure 2). In the scatter plot, the position of each dot encodes the position of the corresponding features of the id's route in high dimensional space. When clicking on a dot, the dot's (route's) features can be shown in texts. The transparency of the dot is 0.5 so that the overlap of dots can be observed. We found that in Figure 1a, there are many dots are highly overlapped, which indicates that many visitors had very similar moving trajectory as others. Previous work [32] also showed that some people's routes are highly correlated, which indicates they move together with their families or friends. Therefore, we performed additional data preprocessing on the two dimensional data that plot Figure 2 to remove the overlapped dots. Briefly, the algorithm is for each small chunk of area (0.1*0.1) in PCA plot, we only visualize 1 ID to represent all IDs in the same area. In this way, we got the left part of Figure 2, which is less dense but preserves most of the properties in PCA plot. The route of each single visitor in the scatter plot can be visualized on the map of Dinofun world (Figure 2 right part). Detailed time, position and movement information can be shown in texts when the mouse is on a specific point on the route. Each record of a visitor is represented by a circle on the corresponding position on the map. The size of the circle encodes the time spent on that position. The color of circles encodes a rough time information: the later, the darker. By exploring the routes of the visitors in PCA scatter plot, the anomalies can be found.

4.2 Place Oriented Approach

We made a Python class "Place" to process the data stored in "Person" objects. Each "Place" object represents either an area or a facility to record the population on that place over time. We visualized the snapshots per 30 minutes of the overall population in the park on the map (Figure 3). The color of the grids encodes the population of each coordinate in the map. Users can press the buttons on the map to change the time of the snapshots taken. If the mouse is over a grid, the detailed information (coordinate and population) is shown. If users click a certain area, the histogram of the population on that area over time is shown on the right.

5. Results

We successfully implemented the people based and place based schemes described in Data Processing session.

5.1 Visitor oriented visualization

Visitor Oriented Visualization is a two component design. On the left side is a PCA scatter plot. We observed that PCA plot (Figure 2) is roughly divided into two clusters. When the number of component = 2, the k-means result can clearly divide the PCA plots into the 2 clusters

that can be easily observed by eyeballing. However, if the number of component > 2, the left dense cluster is further divided into several parts with straight line boundary. The linear division is not natural, so we decided not to apply k-means result to the PCA plot.

The right side graph (Figure 2) is an individual movement visualization. By clicking the dots on the PCA plot, the moving trajectory of the corresponding visitor can be clearly visualized. By tracking the change of the color and the sizes of the circles on Figure 2, we can easily observe the general pattern of a visitor. We found abnormality tends to appeal at the edge of each cluster. For example, the right most dot shows that the visitor with the ID 1781128 only stayed in the park for 3 hours but stayed at the crime scene for 24minutes (Figure 4). Therefore, this design enables the user to clearly track the individual movement and quickly detect abnormal behaviors.

However, the visualization of PCA is not perfect. The PCA results vary significantly based on different features extracted. More hidden anomalies may be discovered by choosing visitor's features to extract more carefully.

5.2 Place oriented visualization

The Place Oriented Visualization is also a two component design. On the left side is a heat map of the whole park. We divided the whole park into 10000 squares based on the coordinator. The population of a square is encoded as color, the lighter the blue, the more population at that time. The whole graph gives a very intuitive view of the overall population distribution. Above the graph are 32 buttons which represents time from 8:00 to 23:00. By clicking those buttons, users can easily have a glimpse on the overall population distribution in the amusement park at a specific time. From Figure 3, we can observe that the most of the populations are distributed on the main road and its nearby facilities.

The bar chart on the right side is the histogram of population over time in a region. There are total five regions which are shown with color blue, pink, yellow, green and red respectively. By clicking any square inside a region, the histogram will show how the total population in the specific region changes. Above the heat map, region name will be shown to indicate which region is chosen. By mousing over the histogram, time and population will be shown.

6. Potential Contribution

6.1 Visitor Oriented Visualization

This visualization can be used to monitoring the park. When there is a kid get lost, someone lose their personal belongings or a crime take place, the park worker can easily get the movement visualization to locate the kid, narrow the places scope or narrow the suspects scope. If there are surveillances inside the park, this visualization can reduce the work of checking the records as it narrows the event inside a smaller range.

6.2 Place Oriented Visualization

This visualization can be used to better assign the personnel and direct visitors. As the heat map gives an overview of population distribution, the park can assign more workers to the crowded region or facility to better serve the visitors. They can also remind the visitors that another rides have a shorter que. In case of emergency such as fire hazard, this system can give an insight of how to evacuate people in order.

The bar chart can provide information that which region is more popular and helps the park to find out the interest of visitors.

In short, our design successfully fulfills the goal of identifying abnormal visitors and it provides a more intuitive demonstration of visitors' behavior which can be used to better manage the park.

References:

1. Aggarwal, Charu C., and Philip S. Yu. "Outlier detection for high dimensional data." ACM Sigmod Record. Vol. 30. No. 2. ACM, 2001.
2. Benjamin, Perakath, et al. "Group identification from visitor movement data: VAST 2015 mini-challenge 1." Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on. IEEE, 2015.
3. Breunig, Markus M., et al. "LOF: identifying density-based local outliers." ACM sigmod record. Vol. 29. No. 2. ACM, 2000.
4. Buchin, Kevin, Bettina Speckmann, and Kevin Verbeek. "Flow map layout via spiral trees." *IEEE transactions on visualization and computer graphics*17.12 (2011): 2536-2544.
5. Buja, Andreas, Dianne Cook, and Deborah F. Swayne. "Interactive high-dimensional data visualization." Journal of computational and graphical statistics 5.1 (1996): 78-99.
6. Cakmak, Eren, et al. "Applying visual analytics to explore and analyze movement data." *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*. IEEE, 2015.
7. N. Cao, C. Shi, S. Lin, J. Lu, Y. R. Lin and C. Y. Lin, "TargetVue: Visual Analysis of Anomalous User Behaviors in Online Communication Systems," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 280-289, Jan. 31 2016.
8. Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM computing surveys (CSUR)* 41.3 (2009): 15.
9. E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection. In Applications of data mining in computer security, pages 77–101. 2002.
10. Forgy, Edward W. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications." *Biometrics* 21 (1965): 768-769.
11. Gra, Denis, et al. "Exploring trajectory data using ComVis CMV tool VAST 2015 Mini-Challenge 1." Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on. IEEE, 2015.
12. Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979): 100-108.
13. Hotelling, Harold. "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology* 24.6 (1933): 417.
14. Inselberg, Alfred, and Bernard Dimsdale. "Parallel coordinates." *Human-Machine Interactive Systems*. Springer US, 1991. 199-233.
15. Kandogan, Eser. "Visualizing multi-dimensional clusters, trends, and outliers using star coordinates." Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001.
16. Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." *IEEE transactions on pattern analysis and machine intelligence* 24.7 (2002): 881-892.
17. Karnick, Pushpak, et al. "Route visualization using detail lenses." *IEEE transactions on visualization and computer graphics* 16.2 (2010): 235-247.
18. Kohonen, T. "The self-organizing map." *Proceedings of the IEEE* 78.9 (1990): 1464-1480.
19. Kruskal, J. B., and M. Wish. "Quantitative applications in the social sciences: Multidimensional scaling (Vol. 11). Beverly Hills." (1978).
20. Laskov, Pavel, et al. "Visualization of anomaly detection using prediction sensitivity." *Sicherheit*. Vol. 2. 2005.
21. LIU, Abishek PURI Dongyu, et al. "ParkVis: A visual analytic system for anomaly detection in DinoFun World." (2015).
22. Lloyd, Stuart. "Least squares quantization in PCM." *IEEE transactions on information theory* 28.2 (1982): 129-137.

23. MacQueen, James. "Some methods for classification and analysis of multivariate observations." *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. No. 14. 1967.
24. Munoz, Alberto, and Jorge Muruzábal. "Self-organizing maps for outlier detection." Neurocomputing 18.1 (1998): 33-60.
25. Munzner, Tamara. "A nested model for visualization design and validation."*IEEE transactions on visualization and computer graphics* 15.6 (2009): 921-928.
26. Phan, Doantam, et al. "Flow map layout." *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*. IEEE, 2005.
27. Richardson, Mark. "Principal component analysis." *URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladn*
28. Steinhaus, Hugo. "Sur la division des corp materiels en parties." *Bull. Acad. Polon. Sci* 1.804 (1956): 801.
29. I. Steinwart, D. R. Hush, and C. Scovel. A classification framework for anomaly detection. In Journal of Machine Learning Research, pages 211–232, 2005.
30. Steptoe, Michael, et al. "VAST Challenge 2015: Grand Challenge-Team VADER/VIS Award for Outstanding Comprehensive Submission." *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*. IEEE, 2015.
31. Wang, Junpeng, Ji Wang, and Chris North. "Spectrum: A visual analytics tool to explore movement logs." Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on. IEEE, 2015.
32. Wei, Shuang, et al. "CrowdAnalyzer: A collaborative visual analytic system."*Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*. IEEE, 2015.
33. Yu, Bowen, and Bo Zhou. "VAST challenge 2015 solver." Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on. IEEE, 2015.
34. Perception in Visualization https://www.csc.ncsu.edu/faculty/healey/PP/
35. http://scikit-learn.org/stable/modules/decomposition.html#pca
36. https://github.com/damjankuznar/pylof
37. http://www.numpy.org/
38. http://jupyter.org/