Visual Analytics of Heterogeneous Data for Criminal Event Analysis

VAST Challenge 2015: Grand Challenge

Junghoon Chae
Guizhen Wang
Benjamin Ahlbrand
Mahesh Babu Gorantla *
Purdue University

Jiawei Zhang
Siqaio Chen
Hanye Xu
Jieqiong Zhao †
Purdue University

William Hatton ‡
United States Air Force Academy

Abish Malik
Sungahn Ko
David S. Ebert §
Purdue University

1 Background

The 2015 VAST Challenge presented researchers with a fictitious scenario that was surrounding a three-day weekend event involving Scott Jones, a renowned football star and a classic hometown hero, at the DinoFun World amusement park. In addition to the regular attractions at the park, the park had arranged for Scott to deliver two speeches on each day of the weekend along with a special exhibit that featured memorabilia in honor of his accomplishments. However, the event was marred by a criminal incident where certain individuals vandalized the exhibition on one of the days of the weekend. The challenge offered two mini-challenges and a grand challenge. The data was comprised of movement and communication information of all the visitors over the weekend. The grand challenge required to combine hypotheses and insights from the heterogeneous data of the mini-challenges in order to analyze and identify Scott's activities, the criminal event, and issues with park operations during the three-day weekend. In this excerpt, we will present our visual analytics framework that was designed for the challenge and our analysis work-flow.

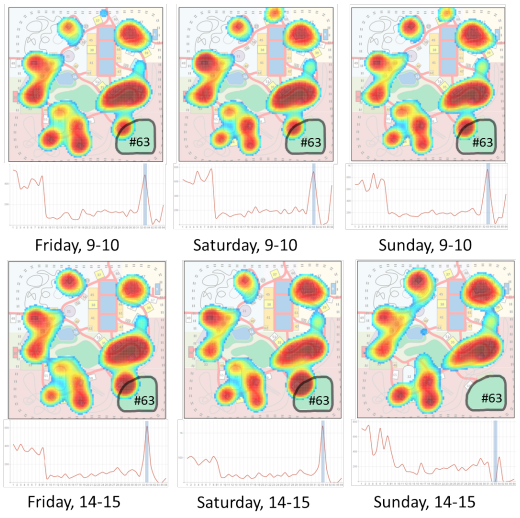

Figure 1: Heatmaps and line graphs based on check-in data.

2 System Overview

Our visual analytics system consists of multiple linked analytic components. Each component is aimed to extract different aspects from the data, such as hotspot extraction, movement tracking, and communication clustering, and is tightly integrated to provide insights into the data. In our system, the movement and communication data are assessed by different analytic techniques and visualized using multiple views. We utilize several clustering techniques to cluster individuals based on different characteristics, including k-means [5] to group people based on trajectory data, longest common sub-sequence [3] and and DBSCAN [4] to cluster individuals with similar attraction check-in sequences, community detection method [1] to cluster visitors based on communication trends, and partition based clustering model [2] to find common sub-trajectories for individuals. We provide a brief background of the challenge data and our methodology below.

2.1 Movement Data

This dataset features movement tracking information of all park visitors. The data includes around 25 million individual movement records with timestamp, visitor id, location coordinates, and type (check-in or movement). Our system allows users to filter and group the movement data, and the results are presented using multiple interactive visual components. We provide a heatmap to extract hotspots for a given time frame. We also build trajectories for each visitor and visualize the trajectories to track their movements. Additionally, we support the discovery of major movement patterns during a specified time period by trajectory clustering.

2.2 Communication Data

The communication data is comprised of about 4 million communication records among the visitors to the park over the three-day weekend. Each communication record includes a timestamp, two visitor ids (for sender and receiver), and two locations (message origin and destination location). To analyze the communication data, we provide communication frequency graphs that show the number of messages transmitted over the three day weekend. These graphs can also be filtered by location. We utilize node-link diagrams to analyze the relationship between groups of visitors, where the edge encodes the strength of relationship between the visitors based on the communication records.

3 Results and Findings

To answer the grand challenge questions, we integrate and build upon the hypotheses and insights we obtained from the mini challenges. The challenge required to generate insights into Scott's activities over the weekend and characterizing the criminal incident that occurred at the park (e.g., when and where the incident occurred, and possible suspects). The challenge also asked to identify issues with park operations during the three-day weekend. In the following, we summarize our analysis results of Scott's activities and the criminal incident.

3.1 Identification of Scott's Activities

In order to analyze Scott's whereabouts over the weekend, we assumed that he attracted a lot of attention from the crowd. We observe that the performance stage became popular at several times

*e-mail: jchae|wang1908|bahlbran|mgorantl@purdue.edu
†e-mail: zhan1486|chen1722|xu193|zhao413@purdue.edu
‡e-mail: C16william.hatton@usafa.edu
§e-mail: amalik|ko|ebertd@purdue.edu

IEEE Conference on Visual Analytics Science and Technology 2015
October 25–30, Chicago, Il, USA
978-1-4673-9783-4/15/$31.00 ©2015 IEEE

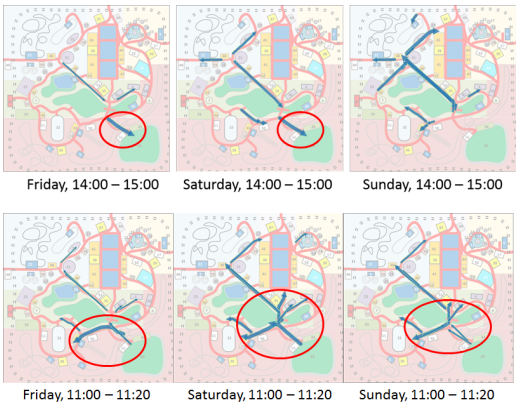

Figure 2: Each figure (top) shows major movements between 2:00 PM and 3:00 PM on each of the three days. Each figure (bottom) presents how visitors move out of the stage after the show.

Figure 3: Communication data from Coaster Alley region in three days. Note the peaks (red boxes) in terms of the number of communications.

during the weekend. We also find that a majority of check-ins occurred later in the hour at around 9:45 AM and 2:45 PM. These results bring us to the conclusion that the performances by Scott started at 10:00 AM and 3:00 PM. Furthermore, the visitors' movements shown in Figure 2 support our hypothesis, where thick arrows indicate how people are moving in before the show began in Figure 1 (Top) and out of the stage area after the show finished in Figure 1 (Bottom). To determine when Scott's performances ended, we noticed a spike of messages sent from the Coaster Alley region an hour after each show started as shown in Figure 3, likely due to people sending messages to other visitors after the end of the show.

While investigating the people who attended the park just to see Scott, we found two individuals who only had one or two check-ins. Our hypothesis was that these individuals would not be visiting the attractions as they were likely going to the park just to see Scott. These two individuals traveled together and took a path we believe to revolve around Scott. We determined that these two individuals arrived at the park and went down straight to the performance stage as shown in the first row of Figure 4. They did this every day, except for Sunday, when they left the park for lunch in the afternoon and did not return, as the second performance of the day was canceled (because of the criminal incident as described in Section 3.2). We hypothesized that these two individuals were either security for Scott or his entourage/friends accompanying him on his path throughout the park each day. The second row in the Figure 4 shows the corresponding movement clusters of the park visitors for the same time periods. We find certain major movements of the public that correspond to the movement patterns of the two individuals; thereby, indicating that the public would flock to see Scott.

3.2 Identification of the Crime

We found that the number of messages sent from the Wet Land spiked unusually from 11:30 AM to 12:00 PM on Sunday. This was likely because the people checking-in to the pavilion after Scott's performance found the vandalism and saw that the medal was stolen. Further, the check-ins to the pavilion ended at 10:00 AM, when the park locked it for the morning show and then restarted at

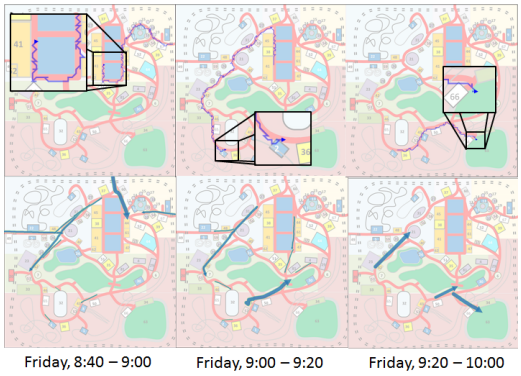

Figure 4: The first row shows Scott's likely movements between 8:40 AM and 10AM. The second row shows the major movements of people for the same time frames.

11:30 AM. However, the last check-in was right before noon, and then after 12:00 PM, no one was allowed to check-in to the pavilion for the rest of the day. This reinforced the idea that the crime was discovered at 11:30 AM. Thus, we concluded that the crime occurred somewhere between 10:00 AM and 11:30 AM on Sunday.

From the news report, we learn that the crime occurred at the pavilion. This was also confirmed by observing that the communications from the pavilion spike on Sunday, and that no one was allowed to check-in there the entire afternoon and evening. Our goal in searching for suspects in the crime was to highlight suspicious behavior of individuals on Sunday. Since we narrowed down the time window and the place of crime, we found that three individuals spent 2.5 hours in the pavilion and were in there alone from 10:00 AM to 11:30 AM. This time frame also matched that of our hypothesis of when the crime occurred. We were thus able to flag these three individuals as potential suspects of the crime.

4 Conclusion

We developed a visual analytics system to analyze the provided heterogeneous 2015 VAST Challenge data. This system utilized several analytic models and visualization techniques. Currently, the underlying data models and clustering techniques have limitations in processing the large volume of data in real time. Therefore, for future work, we will improve the scalability of our system to support real time interactivity and analysis.

Acknowledgements

This work was partially funded by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0006.

References

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[2] J. Chae, Y. Cui, Y. Jang, G. Wang, A. Malik, and D. S. Ebert. Trajectory-based Visual Analytics for Anomalous Human Movement Analysis using Social Media. In *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2015.

[3] C. H. Elzinga. Sequence analysis: Metric representations of categorical time series. *Sociological methods and research*, 2006.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[5] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(1):pp. 100–108, 1979.