Behavior Visual Analysis Supporting Spatial/temporal Data with Uncertainty

Siming Chen[1] Zuchao Wang[1] Xiaoru Yuan[1,2*]

1) Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University
2) Beijing Engineering Technology Research Center of Virtual Simulation and Visualization, Peking University

Abstract

We resent a case study for human behavior visual analytics, dealing with uncertainty in spatial/temporal data. Our data concerns a set of company employees. The data includes their car assignment records, GPS logs for each car and their transaction records. The data are imperfect and there are many errors, data missings and conflicts. With the concept of movement event, we are able to combine these different datasets. By showing an event timeline for each employee, supplemented with a digital map, we can identify five kinds of uncertainties in the dataset. We report two of them in detail.

Keywords: Spatial temporal visualization, Uncertainty, Behavior visualization

1 Introduction

Understanding human behaviors is an important topic, especially in intelligence analysis. In the analysis, We derive behavior information from the spatial temporal data. It poses significant challenges for data processing, because the data is usually heterogeneous and imperfect. There can be data errors, data missing and conflicts. The data can also be in different resolutions. Existing works have defined uncertainty and visualized it in multiple ways [2]. In this paper, we try to identify different uncertainties and refine the them with visual analytics method.

Throughout this work, we use the fictitious datasets from IEEE VAST Challenge 2014 Mini Challenge 2. In our previous work [1], we have introduced a visual analysis system MovementFinder. Based on various visualizations and multiple filters, it is able to summarize the general movement patterns of a group of people, and help analysts detect abnormal events. In fact, many data processing techniques are behind the scene, which help to combine different datasets and deal with various uncertainties. In this work, we combine these different datasets with the concept of movement event. With an event timeline, supplemented with a digital map, we can identify five kinds of uncertainties. We report two of them in detail showing our approach.

2 Data Description

Our dataset is related to a country, Kronos. In the capital, Abila, a big company called GAStech experienced a kidnap. It is suspected that some employees assisted the kidnap, therefore the GPS logs of their cars are provided. The ownership of each car is recorded in a car assignment file. Besides, the transaction logs of the employees are provided, as well as a raster format tourist map of Abila, a vector format road network and a name list of the employees.

The GPS log and transaction datasets cover 54 employees and a time span of two weeks, Jan. 6-19. There are 685,171 GPS records at 1 second resolution. There are two transaction datasets: a loyalty card dataset with 1,391 records, and a credit card dataset with 1,491 records. Most transactions are recorded in both datasets, but they have different temporal resolutions. The credit card dataset records the exact minutes (high resolution), while the loyalty card dataset only record the dates (low resolution).

*e-mail: {csm, zuchao.wang, xiaoru.yuan}@pku.edu.cn

3 Data Fusion

We combine different datasets with the concept of movement event. The events are first defined from the GPS logs, as stops above 1 minute. Each event is naturally associated with a car, a time span and a location. In most cases, an employee is associated with this car in the car assignment data.

After that, we enrich the event data with Point-of-Interests (POIs) and transaction information. For POI enrichment, we first manually extract the public POIs from the tourist map, and put them into many categories. Additionally, we try to identify the home of each employee as the most frequently visited location at 4:00 am. This is treated as a special POI. Then for each event, if its location is within the boundary of a POI, it would have a corresponding label, e.g. "GAStech", "restaurant", "shop", "home". Otherwise it would have a "non POI" label. For transaction enrichment, we first merge the credit card records and loyal card records. Then for each employee, if the transaction time is within the time span of an event, it is assigned to that event.

The fused event data can be visualized with a event timeline (Figure 3-bottom). The timeline can summarize the movement events of one person. The X axis represents hour of day, while the Y axis represents each day. Each event is represented as a rectangle, with the color showing POI categories, and position showing its start and end times. White circle indicates the transaction record, the size of which is the price paid.

4 Uncertainty Processing

During the data fusion steps, we have to deal with uncertainties in various aspects. Our general strategy is to use the event timeline to identify data outliers and conflicts, which often suggests data uncertainties. In many cases, this process also requires some computational methods. After identification, we then try to fix the uncertainties by cross reference between different datasets. In case it fails, we record all possibilities.

For the POI information, it is extracted from the tourist map. However, we only have a rough idea of its position. Therefore, we first tentatively define a polygon region for each POI with Quantum GIS. Then in a later stage, we refined it based on patterns from trajectory analysis. For the temporal information, sometimes GPS logs, credit card records and loyalty card records can give different times for the same event, and we have to guess the most probable one. With the event timeline, we find it is the GPS logs in most cases. Besides, the two transaction data are in different temporal resolutions: one in minutes, another in days. In such case, we use the time with higher resolution (in minutes). For the transaction attributes, the two transaction data can record different person or price for the same purchase. Usually we have no clue of the correct person and price, so we mark both in the event data. Besides, each of the transaction data misses some purchases. By combining them both in the event data, we have a more complete dataset. For the location information, sometimes the GPS logs and credit card records indicate the same person at different locations at the same time. We fix it when possible, otherwise we mark both. The GPS logs can also have data missing, signal shift and noise. This can be observed on the event timeline and digital map. For the people information, there are nine truck drivers without assigned cars, and five cars with assigned drivers. We manually match them by comparing the credit card data of truck drivers and GPS logs of cars on the event timeline. All above uncertainty processings are summarized in Figure 1.

Uncertainty Type	Dealing with Uncertainty	Map	Trajectory Data	Transaction data (low resolution)	Transaction data (high resolution)	Description
POI information	POI region range extraction	√				Extract a range with uncertainty from the JPEG map
	POI region range refinement	√	√			Based on common pattern extracted from the trajectory data
Temporal information	Resolution refinement			√	√	Mark the detail time based on the matched event with high resolution
	Transaction time error and conflict		√	√	√	Mark the conflict time for each transaction events
Transaction attribute	Data missing			√	√	Add and mark the missing records
	Price conflict			√	√	Mark two values
Location information	Location conflict		√	√	√	Mark the conflict location
	Location missing, shift and noise	√	√	√	√	Mark the 'jump' position and shift based on common pattern and POI, using the transaction for verification
People information	Missing car assignment		√	√	√	Find people who have the matched transaction and location of the car

Figure 1: Summary of the five kinds of uncertainties we discovered.

5 Case Study

Due to the page constraints, we mainly focus on two uncertainty processing examples.

5.1 Temporal Error and Mismatch

When we process the transaction data and generate its temporal distribution, we can see several outliers. For example, we find several people with transaction records at exactly 12:00 everyday. As shown in Figure 2-bottomleft, while the credit card data says they are in some restaurant, the GPS logs show that they are not. However, these people usually have been to that restaurant in the morning, but without transaction records at that time. Based on this observation, we conclude that there are temporal errors in the credit card data. To fix such error and get the correct transaction time, we use the time indicated by the GPS logs, and associate the transaction to the movement event in the morning (Figure 2-bottomright). We summarize the general procedure to deal with such errors in Figure 2-top.

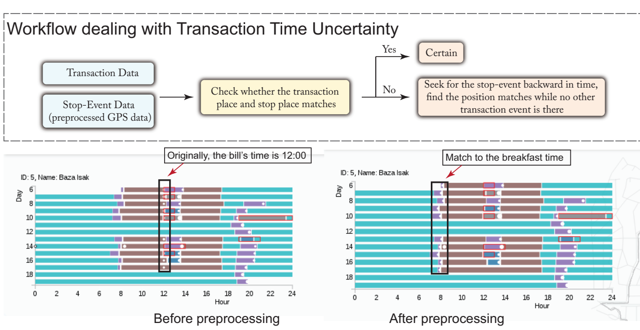

Figure 2: Dealing with transaction time uncertainty.

5.2 Location Data Shift and noise

GPS logs can undergo location shift or noise, but such cases are not easy to detect. With our event timeline, we are able to visualize the temporal distribution of events for each person. In Figure 3-bottomleft, we have identified a person who spends most time in an "Uncertain" POI (yellow), and never goes to the GAStech company (brown). This is strange. By plotting his trajectory on the digital map, as shown in Figure 3-middleleft, we find that his trajectory is highly noisy, and seems having a location shift.

In GPS log processing, the location shift can be regarded as a system error, and the noise as random error. To deal with the noises, we down sample the data. For the location shift, based on the general spatial pattern of other people's trajectories, we identified the shift, and move the trajectory back to the correct location (Figure 3-middleright). Now the POIs on the event timeline are much more reasonable. We find he now spends the working time at the GAStech company, and the locations in GPS logs matches perfectly with that in the transaction records (Figure 3-bottomright). We summarize the general procedure to deal with such errors in Figure 3-top.

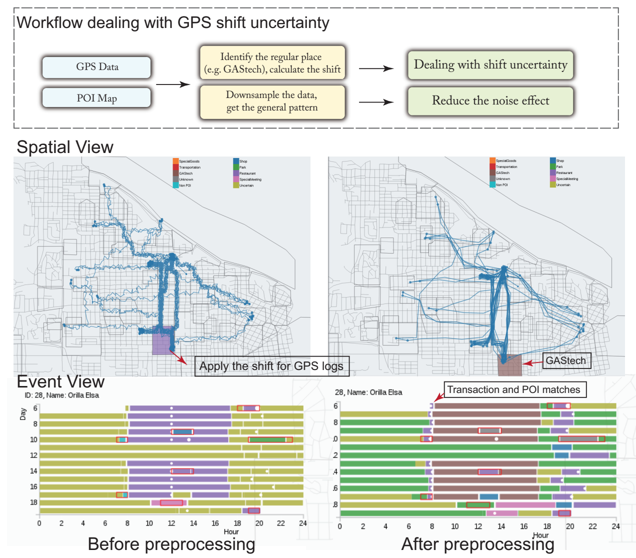

Figure 3: Dealing with shift and noise in the location information.

6 Conclusion

In this paper, we have presented a case study for human behavior from spatial temporal data. Our research emphasizes on detecting and handling the uncertainties in heterogeneous dataset, including car assignment data, GPS logs and two transaction data. We combine these different datasets with the concept of movement event, and visual them in a event timeline. With a fictitious dataset, we have identified five kinds of uncertainties, and we detailed two of them. This shows the flexibility and usefulness of our method. In the future, we would try to handle larger scale of data. Besides, we would try to generalize the uncertainty defining and processing solution, for extending to other types of data and complex situation.

Acknowledgements

The authors wish to thank reviewers and IEEE VAST Challenge committee. The authors wish to thank Zhenhuang Wang, Chenglong Wang, Zipeng Liu and Zhengjie Miao for contributing to this work. This work is supported by NSFC No. 61170204.

References

[1] S. Chen, Z. Wang, Z. Liu, Z. Wang, C. Wang, Z. Miao, and X. Yuan. Movementfinder: Multiple filters for spatial temporal visual analytics. In *Proc. of IEEE VAST, Poster*, 2014.

[2] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005.