Visual Analytics for Detecting Behaviour Patterns in Geo-Temporal Data

Michael Hundt* University of Konstanz
Natascha M. Siirak† University of Konstanz
Manuel Wildner‡ University of Konstanz

ABSTRACT

Today raw data get more complex and are obtained from different sources. In this paper, the data sources evaluated are GPS tracks of thirty five cards and purchase records of fifty four persons over a time span of two weeks. Additionally, a map and more background information has been provided. Visual Analytics is indispensable for event- and pattern-recognition and the understanding of as well as getting a sense for spatial properties. This paper describes a journey from dealing with geo-temporal data over creating a data foundation for further means to an exploration tool that enables every user to access the data in an interactive, fast and non-exhausting way.

Index Terms: H.1.2 [User/Machine Systems]: Human information processing—; H.2.8 [Database Applications]: Data Mining—Spatial databases and GIS; H.3.3 [Information Search and Retrieval]: Information filtering—

1 INTRODUCTION

The VAST Challenge of the year 2014 is about identifying suspects and possible victims and detecting underlying circumstances of an abduction. The scenario is set up on the fictitious island of Kronos where the affected company GasTech is located. This paper deals with the mini challenge 2, which is about the determination of the normal behaviour of the typical employee of GasTech, the identification of unusual behaviour and events and an overview of our data-handling. We provide the analytical methods as well as the visualization techniques that were applied to detect behaviour patterns in geo-temporal data. The available data sources are GPS-tracks of company cars, a table containing the assignments of cars to certain people, credit-card data and loyalty-card data of employees and other persons. Further a simplified map of Abila, the capital of Kronos is given. The data cover 14 days, from 6th Jan, 2014 to 19th Jan, 2014. Limitations of existing tools for analysis and visualization of the data made it necessary to develop a custom visual analytics tool.

2 DATA PROCESSING

The data have been processed iteratively by applying preprocessing, visualization. Also, new data have been derived as a result of that process. Visualization is considered an expedient and important step that gives impulses for further processing. In the end we are able to create a geo-spatial database that contains our processed and derived data for easier integration into the tool. This database provides easy access to state-of-the-art geo-spatial analysis.

2.1 Preprocessing

Preprocessing is mainly done by using KNIME. Most of the steps are integrated into the workflows that are used to create derived data. This facilitates a flexible handling of the preprocessing. The data contain missing values, e.g. unavailable car-assignments for truck drivers, and erroneous data, such as bookings of purchases at an implausible time. It is also necessary to perform transformations, for example conversion of strings to dates. Some of these procedures are already covered by existing nodes in KNIME, e.g. string-conversion. Others require a more sophisticated approach. For example, the wrong booking times are substituted by the times the purchases most probably took place, during a matching car-stop. Sometimes an automatic correction is not possible and we have to check if either the data can be ignored or special considerations are needed.

2.2 Derived Data Creation

The necessity to create derived data to gain more complex insights is revealed by looking at first visualizations, like scatter plots in KNIME and GPS-data in QGis. A KNIME-workflow is used to reduce the GPS-data. It extracts the coordinates where cars stop for more than 30 seconds and creates "stop-locations". Based on these, "home-locations" are created at the places where people most probably live. These are the locations where persons spend most of their nights. The approximate locations of shops and the GasTech headquarters are determined to facilitate the comparison of the whereabouts of persons. This is achieved by joining the credit-card- and gps-data. The results are quite secure regarding the shops. The locations where people did not buy anything yield less reliable results as there is no further confirmation of the stops. During this process some stop-locations stand out: several people stop there, there are no related purchases and they are not in a reasonable distance to any location on the tourist-map. These "unknown locations" represent unusual behaviour.
Furthermore locations are categorized by usage: defined by attributes like average money spent, number of visits, time of purchases/stops etc. Another category of derived data is required by our tool, for example, parks are vectorized to be able to show them as landmarks on the map.

3 EXISTING TOOLS

3.1 Early Insights

By utilizing KNIME and QGis we were able to solve parts of the tasks in mini challenge 2. We got an overview of the movements of single persons, e.g. movements late at night, and of the locations where people meet. It is possible to compare certain attributes one by one. Furthermore QGis allowed us to vectorize data and create new layers, e.g. to display home-locations. It also allows to create visualizations like heat maps.

3.2 Limitations

Using existing tools we experienced several limitations. For example, the state of the system is not always visible, for instance, the current setting of filters is not displayed directly on the GUI in QGis. Free exploration is not possible: the setting of filters is complex and everything has to be entered manually. The user has to edit several filters to perform complex queries like "where were these four people on that day and at that time". All in all the existing systems feature high interaction costs, low reproducibility, high interpretation costs and high goal forming cost[1]. Large interaction

*e-mail: michael.hundt@uni-konstanz.de
†e-mail:natascha.siirak@uni-konstanz.de
‡e-mail:manuel.wildner@uni-konstanz.de

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

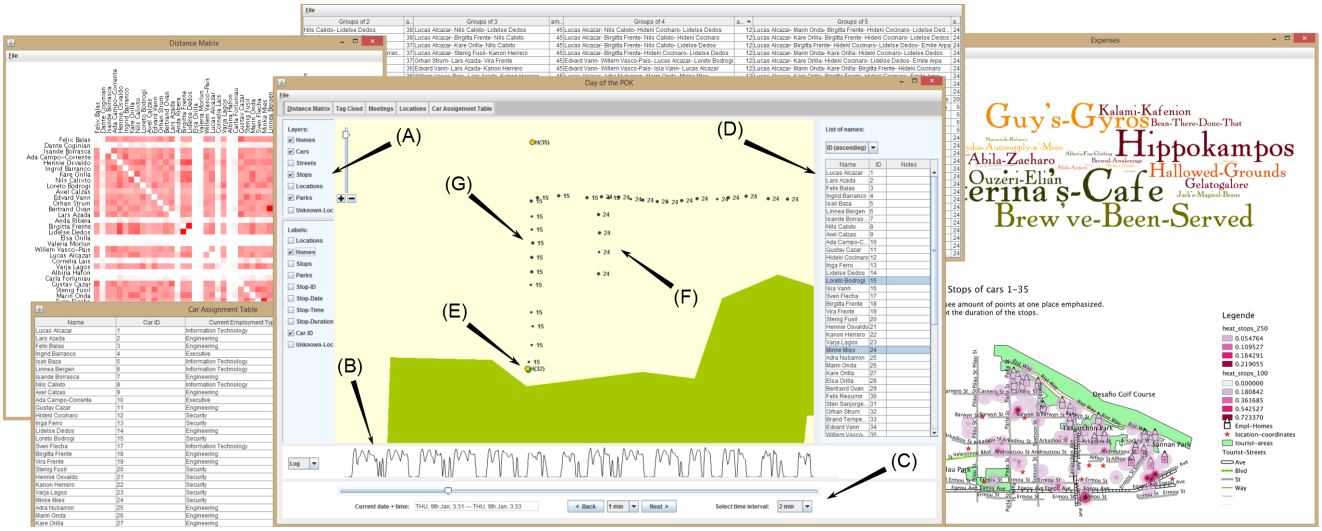

Figure 1: This picture gives an overview of the possibilities our tool offers. You can see two security members take turns in visiting the home of an executive in the middle of the night, guarding or tailing? (A) Layers & Labels, (B) Car Movement as Line Chart, (C) Time Slider & Options, (D) Employee Table, (E) Home of ID 32, (F) Car 24 Arriving, (G) Car 15 Departing

costs weary the user. In the worst case he loses his target and is frustrated, stops his exploration and does not gain further insights. We defined our specific needs and designed a tool that allows a more intuitive exploration of the data to counteract these limitations.

4 Overcoming Limitations by Designing a New Tool

4.1 Solutions

The primary goal of our design is to facilitate interactive exploration of the data. In the center of the window is a map that shows the city and parks. Several layers are only displayed if selected to avoid clutter. Filtering for persons is simply done by selecting them in the table on the right side of the map. Stops and GPS-data will be filtered and displayed accordingly. A slider enables the user to look at a certain point in time. It is also possible to select the time span that is shown as well as jumping through the data in predefined steps, e.g. several minutes or hours. Changes of the range of and the point in time directly affect the displayed data. A line-chart that shows an aggregation of the movements of all cars helps to find interesting events. Further incentives to explore are given by predefined sortings of the person-table (e.g. by total movement) and visualizations like a heat-map (Figure 1). The state of the system is always visible: represented by the highlighting of persons, the position of the time-slider, a label that shows current date, time and timespan, and the display of the sorting that was chosen. This composition of features addresses most of the limitations we mentioned previously in 3.2.

4.2 Evaluation

Our tool minimizes interaction costs by drastically reducing the complexity of filtering. Thereby it encourages free exploration and enables the user to find previously undiscovered events or errors in the data. For example, we realized by seeing a "jumping" GPS-data-point, where a car started at a different location as it stopped, that there is missing data and not a long stop. Aided by our program it is possible to compare the movements of different persons in an easy way. The control over time and the simple filtering and selecting of persons enables detailed observations. Figure 1 shows a time-crucial and arranged event where probably two security members tail an executive at his home. The tool is designed for interactive exploration of the data, especially of spatio-temporal data. It does not frustrate the user early on, so exploration will not be terminated prematurely [2] [3].

4.3 Future Work

There are several means to enhance our approach. We propose the implementation of semantic zooming to improve access to detailed information. An example would be an overview of the spending of persons that are at a certain location when the user selects that location. Another improvement would be to enable the user to save the state of the program, e.g. the time-slider position and all the selections he did, to facilitate the reproduction of this state and thereby the refinding of an event. We also propose more possibilities for advanced analyses, e.g. of the spending habits of persons, by further visualizations that can be accessed via the tool.

5 Conclusion

In this paper we show our approach to the solution of the tasks. We explain our usage of existing tools and their importance for our initial steps. Those steps are described as well as the limitations we met. Out of these limitations emerges the need for a tool, that facilitates data exploration to enable a user to gain more complicated insights. We describe our approach that derives from the necessity to reduce interaction costs. The ensuing evaluation shows that our tool enables the user to gain more insights in a less tiring way. We propose means to further enhance our approach.

References

[1] H. Lam. A framework of interaction costs in information visualization. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1149–1156, 2008.

[2] T. Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, Nov. 2009.

[3] M. Sedlmair, M. Meyer, and T. Munzner. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis)*, 18(12):2431–2440, 2012.