ODIX: A Rapid Hypotheses Testing System for Origin-Destination Data

PREPRINT

Juri Buchmüller* Universität Konstanz
Wolfgang Jentner† Universität Konstanz
Dirk Streeb‡ Universität Konstanz
Daniel A. Keim§ Universität Konstanz

ABSTRACT

In this paper, we present our solution to the VAST Challenge 2017 Mini Challenge 1. We discuss challenges posed by data set and tasks and introduce ODIX, a custom rapid hypotheses testing system tailored to origin-destination data as provided by the challenge. We show findings made with ODIX and illustrate how we apply sequential pattern mining to explore common traffic patterns.

Index Terms: H.4.0 [Information Systems]: Information Systems Applications—General

1 INTRODUCTION

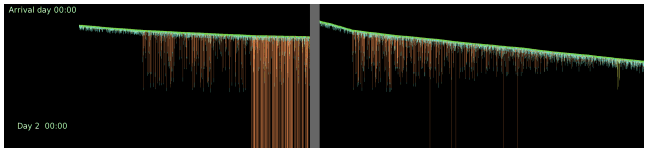

Figure 1: Day tourists arrive earlier than over-night campers (left part). The earlier they arrive, the longer they stay in the evening (right part). Each column represents one car. Columns are aligned by the time of day and ordered by the time they enter the reserve. Colors are mapped to types of places.

The setting of the 2017 VAST Challenge is a fictitious nature preserve, which is developed with a road network. In its vicinity the town of Mistford is home to some industry. A local ornithologist observes an alarming decrease in the population of a bird species. In Mini Challenge 1 (MC1), the task is to find possible reasons for this decrease by characterizing the motor traffic within the preserve and detecting unusual behavior.

The given traffic data set comprises Origin-Destination data (OD) for seven categories of vehicles for thirteen months, with the events taken at neuralgic points of the preserve, such as entrances, campgrounds or traffic gates. The specific tasks can be split into two categories requiring different analysis methods: First, spatiotemporal patterns need to be identified in both short and long-term scope, requiring aggregated views on the data in parameterizable intervals. Second, the identification of unusual events and temporal patterns with few participating vehicles requires an individual view on single trips. Together, findings of both types have to be analyzed for their potential impact on bird life.

In a first step, after a thorough examination of the given data using R [2], we came up with a list of coherences and hypotheses that would have to be checked in order to answer the tasks. For example, we characterized traffic for each vehicle category spatially, temporally, in combination, by road restrictions or by speed thresholds. As the diversity of conditions to consider is very broad, a fully automatic approach to extract patterns and outliers would be inappropriate. Furthermore, looking at the data, a user quickly generates hypotheses he wants to check. For this reason, we decided to implement an own solution that helps a user to rapidly check hypotheses by efficiently combining categorical, statistical, spatial and temporal filter options in one place. The system further assists the user by automatically extracting sequential patterns based on filtered selections, allowing the user to determine typical traffic patterns. For further pattern analysis, we reused parts of the Visual Movement Explorer [3] programmed for the VAST Challenge in 2015, with only slight modifications and some additional data preparation.

In the remainder of this paper, we discuss assumptions and heuristics applied in preprocessing steps and provide a description of our tool. We also present some findings and illustrate some traffic patterns we extracted.

*e-mail: buchmueller@dbvis.inf.uni-konstanz.de
†e-mail: jentner@dbvis.inf.uni-konstanz.de
‡e-mail: streeb@dbvis.inf.uni-konstanz.de
§e-mail: keim@dbvis.inf.uni-konstanz.de

2 DATA PREPROCESSING

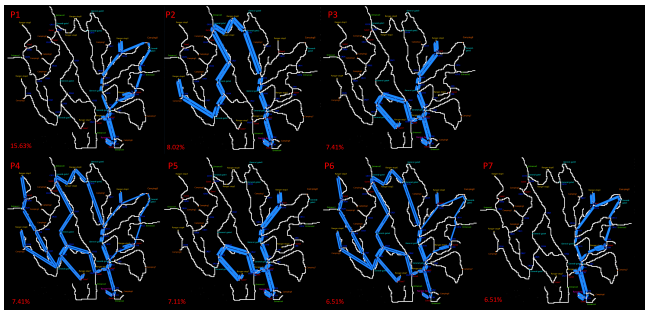

Figure 2: Typical patterns of ranger patrols derived using frequent pattern extraction, ordered by their share amongst all ranger patrols. The first and the last as well as the third and fifth pattern feature almost the same route, but patrolled in opposite directions.

We start by aggregating the data into individual sequences. Each sequence represents one trip. The car-type is a metadata property of a sequence. The sequence consists of items which are defined by the timestamp and the location of the car (e.g. gate8, camping1, ...). We recognized that the car-ids are reused sometimes. Therefore, we applied a split heuristic: We split trips as soon as a car leaves the park or arrives at a ranger station. We appended a trip-suffix to the existing trip-id to create a unique id for each trip.

The extraction of trips enabled us to derive further statistics about trip length, stay time at points of interest such as campgrounds, and speed of the vehicles. Since the sensor network is sparse and does not cover every road crossing, ambiguities in route options between two recorded positions arose and had to be resolved. To do so, we extracted the coordinates of all sensor locations from the map given and added new helper positions at each road crossing (dark blue in Fig. 3C). This also helped to achieve a more accurate approximation of distances between two recorded points. We then modeled the road network as a graph in the graph database neo4j[1] and implemented individual edge cost functions to reflect distances

[1]https://neo4j.com/, accessed 11.08.2017

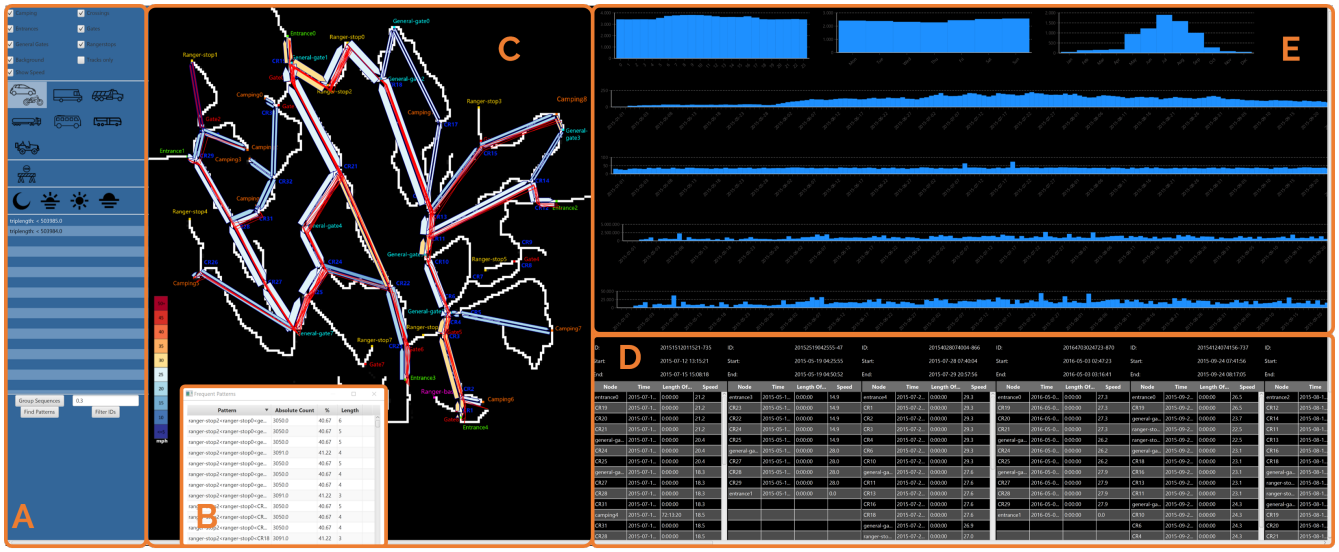

Figure 3: The ODIX interface: Filtering and Sequence Analysis options are available in the left sidebar (A) and detailed information about extracted sequences in the sequence detail window (B). The main map view (C) shows the amount of traffic and average speed for street segments as well as individual trips. Bar charts on the right (E) provide temporal distribution in semantic aggregation levels as well as the trip amount, speed, duration and length statistics. Single trips can be analyzed in the detail view in (D).

and roads restricted to ranger traffic. For all trips, we assumed that vehicles take the shortest route between two recorded positions. Using the Cypher query language of neo4j, we were able to calculate individual vehicle speed per segment as well as aggregated vehicle speeds. Also, road restriction violations can be queried easily.

3 Application

For the first glance at traffic patterns, we reused a tool developed for the 2015 VAST Challenge: With the Visual Movement Explorer [3] we were able to investigate behavioral patterns over time. For example, day tourists arrive earlier than tourists that stay over night (see Fig. 1). The yellow peak on the right is even more interesting, which we identified as some special event at a ranger-stop after a more detailed investigation.

Yet, not all questions could be answered using available tools, so we chose to implement ODIX, a novel prototype helping to explore OD data both in detail and in aggregated form. The ODIX interface consists of three main categories: The left side (Figure 3, A) shows general settings and categorical filters, e.g. car types or violations. The center(C) gives the user spatial access to the data where the underlying graph structure is mapped onto the map of the park. The right side (E) provides temporal information. The upper three bar charts aggregate the number of trips per hours, weekdays, and months. The bar chart below shows the number of trips for each date. The user can set a date-range filter by clicking on this chart. The lower three bar charts use statistical data: speed, stop time, the length of a trip. For each, the user can either select to display the maximum or the average per day. Additionally, the user can add filters based on these statistics. For example, the user can filter trips with higher or lower speed than a given value. By clicking on one of the segments in the map, a spatial filter is applied to filter trips that pass through this segment also considering the direction. Finally, it is possible to group sequences. This presents the user all distinct trips in the current filter selection. Several trips can be selected to further filter down the data. We used this strategy to find and generate the typical ranger patterns as depicted in Figure 2.

Similarly, the user can search for frequent sequential patterns. These are also displayed in a table and can be used as a filter. All filters that are currently in use are listed on the left side. Each of the filters can be negated or removed individually. Not shown is a table, which is located below, displaying each individual trip. Selecting trips there highlights their route in red in the graph visualization as well as opens a detail pane for each trip (D). This helps the user to investigate a few selected trips in detail and was used, for example, for detailed comparisons of cars involved in a race.

4 Conclusion

The seemingly fairly simple data set turned out to be difficult to analyze automatically, especially concerning the broad spectrum of assigned tasks. Yet, we quickly formed hypotheses about how to solve the tasks and, thus, concentrated on building a system that allowed to rapidly test these hypotheses visually, assisting the user with comprehensive filtering options. With the sequential pattern mining, we offered automatic support where human cognition comes to a limit. In this case at regular pattern extraction over longer periods of time. As well, many of our findings were also made in a collaborative effort, where three analysts used ODIX in a large screen environment. This can be seen in our video [1].

To improve ODIX, we want to extract general criteria for interestingness both in regular OD patterns as well as in unusual occurrences, and to leverage automatic detection methods based on both measures.

Acknowledgments

This work was supported by the EU project VALCRI under grant number FP7-SEC-2013-608142.

References

[1] J. Buchmüller, W. Jentner, and D. Streeb. Vast challenge 2017 mc1 - team birds and recreation. `https://www.youtube.com/watch?v=il1qtCqf_0M`, 2017.

[2] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[3] D. Streeb, U. Schlegel, J. Buchmüller, F. Fischer, and D. A. Keim. Using Visual Analytics to Analyze Movement and Action Patterns. In *IEEE Conference on Visual Analytics Science and Technology (VAST Challenge 2015 MC1)*, 2015. doi: 10.1109/VAST.2015.7347665