Eventpad: Rapid Event Mining with Visual Analytics

Bram C.M. Cappers and Jarke J. van Wijk

Abstract—In this paper we demonstrate how we can significantly speed up the discovery of anomalies and patterns in event sequences using our visual analytics system Eventpad: the notepad editor for event data. We illustrate the effectiveness of multivariate regular expressions, pattern aggregations, and selections to quickly discover patterns within and between sequences. In particular we show how we used Eventpad to win the IEEE Visual Analytics Science and Technology 2017 Challenge 1 by reducing months of data analysis to just a few hours.

Index Terms—Visual Analytics, Event Visualization, Pattern Mining, Regular Expressions.

1 PURPOSE
Companies try to protect and improve their systems by recording their behavior in logs. Telecom companies for instance analyze their communication networks for the presence of fraud, hospitals analyze patient treatments to discover bottlenecks in the process, and companies study their workflous to improve customer satisfaction. The common ground here is that domains are interested in the analysis of sequences (e.g., phone calls, treatments, workflows) by recording events.

Analyzing event collections is still a challenge due to size and variety. Besides their type and timestamp, events often have additional attributes depending on their domain (e.g., source, length, status information etc.). Current methodologies often limit their analysis to a single attribute to keep the use of (automated) techniques computationally manageable. For root-cause analysis of anomalous sequence however it is crucial that all information is taken into account.

To enable quick and effective analysis of attributes and event sequences, we designed a visualization system where users can visually explore event collections, create rules for patterns they are interested in and let artificial intelligence techniques automatically discover new patterns according these rules. To support this flexibility we introduce Eventpad: the notepad editor for event data. More specifically, our main contributions are:

- a visual analytics system for quick discovery and identification of patterns in large event sequence collections.
- a visual query interface to find and replace event sequences by higher level concepts using multivariate regular expressions.

We refer to the work of Cappers et al. [2] for more details.

2 METHODS
Eventpad is a visual analytics system that achieve rapid pattern discovery by tightly combining visualization, process mining, and automated techniques in one interface. Similar to notepad editors, the system uses find and replace functionality to discover patterns inside the data. To study commonalities and differences between such sequences, Eventpad enables users to simultaneously explore sequential patterns alongside their attributes.

In Eventpad event sequences are visualize as series of “blocks” (also depicted in Figure 1). Initially, all block are colored Gray. Users can create rules to replace block sequences by a new (custom styled) one based on event properties that are of interest. This enables user to incrementally rewrite data and visually encode (sequential) properties that are relevant for investigation. Rules can be used for three purposes, namely to:

- discover new patterns by visually encoding properties of interest;
- test the presence of sequences through pattern matching; and
- remove sequences that are obsolete for the investigation.

Similar to regular expressions, operators such as sequential composition, iteration (0 or more times), and choice are used to construct more complex queries. Automated techniques such as clustering and alignment in turn can use the visual encodings to discover patterns between block sequences. Clustering enables users to study pattern frequencies, whereas alignment can detect overlap between similar sequences using Multiple Sequence Alignment [1]. Figure 1 shows a schematic overview of the main workflow in Eventpad.

- Bram C.M. Cappers and Jarke J. van Wijk are with Eindhoven University of Technology. E-mail: b.c.m.cappers@tue.nl and j.j.v.wijk@tue.nl.
3 RESULTS

Eventpad was used in the IEEE Visual Analytics Science and Technology Data Challenge 2017 to study vehicle travel patterns inside a nature preserve. Competitor teams from both Academia and Industry on average worked 2 months on the data to solve the challenge. With Eventpad we managed to do this in 2 hours. For this we received the IEEE Visual Analytics and Science award for “Elegant System for Hypothesis Testing and Generation”. For a better experience of the interaction and other use cases in practice, we refer to the supplementary video 1 and website 2.

3.1 Exploration

The Lekagul Natural Preserve stores sensor events for every vehicle driving through gates inside the habitat. The data set consists of approximately 100,000 events and over 10,000 vehicles. Besides a timestamp and car-id these events store additional multivariate data such the type of vehicle it corresponds to and the gate name it passed.

We initially start the search for daily patterns by grouping the data by car-id per day. We construct five rules where we color camping events orange, entrances in green, general-gates in blue, rangerstops in yellow, and rangerbase events in pink. For the inspection of frequent daily patterns in the data, we cluster sequences based on their visual representation and sort them by frequency (Figure 2A). Applying Multiple Sequence Alignment on the most frequent sequences enables us to identify four main patterns, namely vehicles entering, leaving, and driving through the preserve along with ranger traffic (Figure 2B).

To study enter and leave behavior of vehicle types, we group the traffic by car-id only. We construct a rule that compresses all enter and exit patterns of vehicles in the data into purple blocks (Figure 4A). We do this by stating that in between two entrance events, no other entrance events are allowed (Figure 4B). This reveals that certain vehicles visit the preserve multiple times in a year (Figure 4A). Selecting the long sequence in Figure 4A and disabling the constructed rule shows that this sequence corresponds to a 2axle truck driving systematically between entrance 4 and camping 4 during high season only (Figure 4C).

According to the challenge description, only ranger vehicles are allowed to travel through rangerstops. In Eventpad we can easily verify this statement by searching for sequences with rangerstops whose car type differs from ranger vehicles (Figure 3A). This reveals 23 cases where 4axle trucks are driving midnight between rangerstop 3 and entrance 3 (Figure 3C). Inspection of the event properties in these sequences in a tabular view shows that these vehicles only drive on Tuesdays and Thursdays.

4 CONCLUSION

The Eventpad prototype demonstrates how combined visual exploration of event collections and sequential analysis can be used for quick and efficient anomaly detection. The ability to visually encode parts of the data based on rules enables analysts to incrementally label their traffic and define their notion of what desired and undesired behavior looks like. The labeling in turn can be used by automated techniques to reduce the solution space and provide better context sensitive insights.

ACKNOWLEDGMENTS

This work is funded by SpySpot, a project in the Cyber Security program of Netherlands Organisation for Scientific Research (NWO).

REFERENCES
