**Detecting Patterns in Event Sequences: Entropy, Acceleration, Clusters & Gaps**

Spring 2019 Directed Study Project (January 5, 2019), University of British Columbia

Student researchers: Jamie Dunkle [jdunkle@psych.ubc.ca](mailto:jdunkle@psych.ubc.ca)

Darius Muglich [dariusm1997@yahoo.com](mailto:dariusm1997@yahoo.com)

Faculty advisors: Malabika Pramanik [malabika@math.ubc.ca](mailto:malabika@math.ubc.ca)

Ben Shneiderman [ben@cs.umd.edu](mailto:ben@cs.umd.edu)

**Abstract:**

There are a rich variety of patterns in time-stamped event sequences such as in medical histories (heart beats, medication, etc.), exercise regimes (dates of workouts), Internet of Things logs (door open events), social media streams (time stamps for tweets from an account or uses of a hashtag), earthquake aftershocks, etc. The simplified case, which is our initial focus has only one event type and only time stamps, not even values for the events. Later we will address more complex cases.

We begin with small synthetic data sets so that we can test our algorithms that detect meaningful patterns, plus visually present them to support human understanding. The first small synthetic data set (20 sequences with 20 events) will exercise our algorithms starting with metrics for detecting:

(1) entropy (regular/irregular patterns, uniformity),

(2) accelerating/decelerating event frequencies, and

(3) clusters/gaps, as well as dense/sparse regions

We seek to algorithmically rank them for the degree of regularity, frequency of change, or clusterness/gapness. We will also develop visualizations to allow inspection of these event sequences, so as to verify the correctness of the algorithms and enable analysts to spot anomalies that could lead to improved and new metrics

1. **Introduction**

<<Maybe Malabika could improve this>>

Time-stamped event sequences are a rapidly growing research topic because of the voluminous data produced from electronic health records, self reports of diet and exercise, Internet of Things logs, social media streams, earthquake aftershocks, and much more. Our goal is develop algorithms and visualizations to enable analysts to better understand the patterns in these event sequences such as (1) regular events such as daily medications or heart beats, (2) accelerating/decelerating event frequencies, and (3) clusters/gaps, as might occur for social media postings. We seek to algorithmically rank the event sequences by their degree of regularity, frequency of change, or clusterness/gapness.

We will explore how to display an event sequence, so a person (including us) could see the pattern that the algorithm finds? In our earlier work we used triangles for events in LifeLines2 and then EventFlow:

<http://www.cs.umd.edu/hcil/lifelines2/>

<http://hcil.umd.edu/eventflow/>

We are reaching out to others at UBC and beyond to learn what they are doing, ask them for test data sets tied to realistic problems, and let them know what we are doing.

Eventually we will deal with

1. much longer event sequences (millions),
2. event sequences where the pattern changes over time (regular for a while, then chaotic, and then back to regular), and
3. event sequences that have multiple event types (as long as a patient takes their medication, their heart rate is regular, but within 36 hours of stopping medication they have abnormal heartbeat events.

We think there are opportunities for applied and basic research, as well as software implementation for automatic detection and user interfaces to support exploration and visual display of event sequences.

1. **Previous & Related Work**

<<who’s ready to start writing this? If Simon is joining the group, maybe he could take this task>>

1. **Test data set construction**

<<Jamie could write this>>

What are the design constraints for the first small test data set:

* 20 cases, each with 20 events
* Ascending order of event times, no two events at the same time
* Maybe limit time between events to the range of 1 to 1000 seconds?
* Maybe limit the range of times from first to last event to 10,000 seconds?
* The first data set should contain cases that show only one pattern, avoiding a complexity such as an accelerating sub-sequence followed by a decelerating sub-sequence.

<< could Jamie express these constraints in mathematical form?>>

Maybe the second medium test data set would have 100 cases with 100 events in each (10,000 events total) and the third large test data set would have 1000 cases with 1000 events in each (1,000,000 events total).

As we move on the more realistic test data sets will have different numbers of events in each case.

1. **Design of event sequence visualizations**

<<Jamie could write this>>

This would cover screen design with great figures of comprehensible images with legends, labeled axes and clear captions to include in the paper. We agree that the visualizations would align on the first event. The question is whether the last events will be aligned, or if the event sequence visualizations would be of different lengths. As we go forward dealing with more events in each visualization will require some creative solutions.

1. **Algorithms and metrics**

<<Darius could write this>>

This would cover the design of metrics of metrics for entropy, acceleration, and probably several metrics for clusters/gaps. The metrics could be [0, 1] or [0, infinity], but the deceleration metric might be negative.

Cluster/gap metrics will need to deal with the number of clusters in a test case, relative size of a cluster (e.g. 6 of 20 events) and its tightness (characterization of how isolated it is from other events). In short we need a clear definition of what is a cluster/gap.

1. **Applications**

<<maybe Malabika could reach out to some colleagues>>

* We need to explore: medical histories, twitter streams, seismic activity, IoT logs, etc.
* Who at UBC is working on these kinds of problems?
* Working on real applications will greatly strengthen our work

1. **Results and Discussion**
2. **Future Directions**
3. **Conclusions**

**References**

Chazelle, B., Drysdale, R. L., & Lee, D. T. (1986). Computing the largest empty rectangle. *SIAM Journal on Computing*, *15*(1), 300-315.

Edmonds, J., Gryz, J., Liang, D., & Miller, R. J. (2003). Mining for empty spaces in large data sets. *Theoretical Computer Science*, *296*(3), 435-452.

Gotz, D., Wang, F., & Perer, A. (2014). A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of biomedical informatics*, *48*, 148-159.

Guo, S., Jin, Z., Gotz, D., Du, F., Zha, H., & Cao, N. (2019). Visual Progression Analysis of Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 417-426.

Nguyen, P. H., Turkay, C., Andrienko, G., Andrienko, N., Thonnard, O., & Zouaoui, J. (2018). Understanding user behaviour through action sequences: from the usual to the unusual. *IEEE transactions on visualization and computer graphics*.

Sadahiro, Y., & Takami, K. (2001). A statistical method for analyzing the relative location of points in a bounded region. *Geographical Analysis*, *33*(1), 40-57.

Seo, J., & Shneiderman, B. (2005). A rank-by-feature framework for interactive exploration of multidimensional data. *Information visualization*, *4*(2), 96-113.

Shneiderman, B. (2016). The event quartet: How visual analytics works for temporal data. In *IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis. IEEE*.