VIT-PLA: Visual Interactive Tool for Process Log Analysis

Sen Yang[1], Xin Dong[1], Moliang Zhou[1], Xinyu Li[1], Shuhong Chen[1], Rachel Webman[2], Aleksandra Sarcevic[3], Ivan Marsic[1], Randall Burd [2]

[1]Rutgers University, Piscataway, NJ

[2]Children's Nat'l Medical Center, Washington, DC

[3]Drexel University, Philadelphia, PA

Abstract

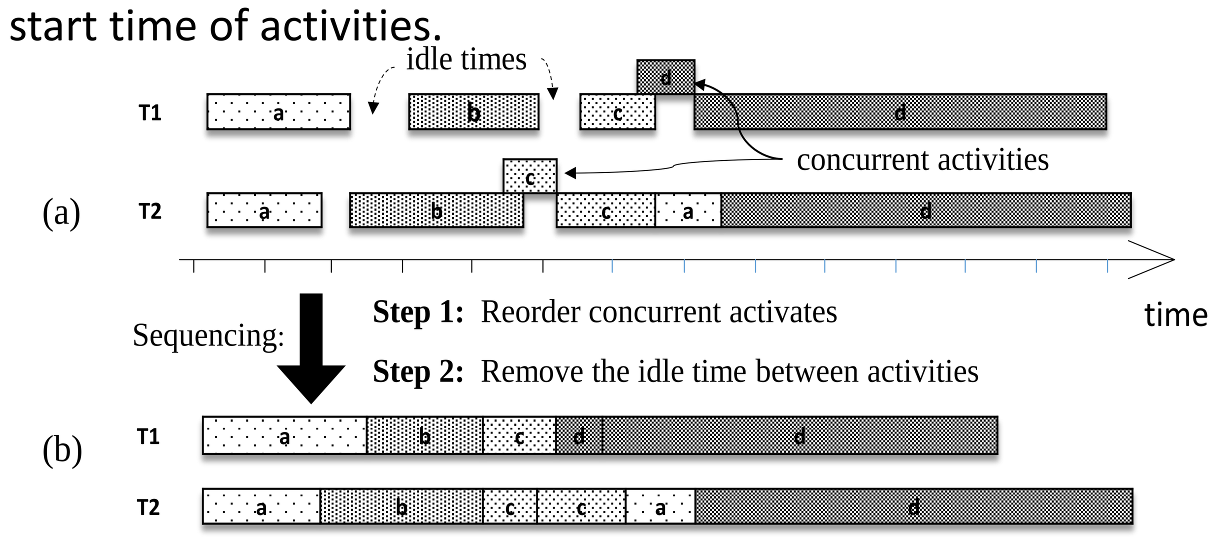
Techniques for analyzing and visualizing process or workflow data have been developed and applied in a wide range of domains (Van Der Aalst, W. et al. 2011; Monroe, M. et al. 2013). Visual analysis of large process logs and integration of statistical analysis, however, have been limited. We introduce the Visual Interactive Tool for Process Log Analysis (VIT-PLA) that provides a simplified process log visualization and performs statistical correlation analysis on process attributes. We demonstrate its use by applying it to an artificial dataset and running a preliminary analysis of trauma team task data collected from a medical emergency department.

Objective

1. Simplify visualization of large process logs (workflow data)
2. Uncover the correlation between process trace clusters and process trace attributes

Data Preprocessing: Sequencing of Traces

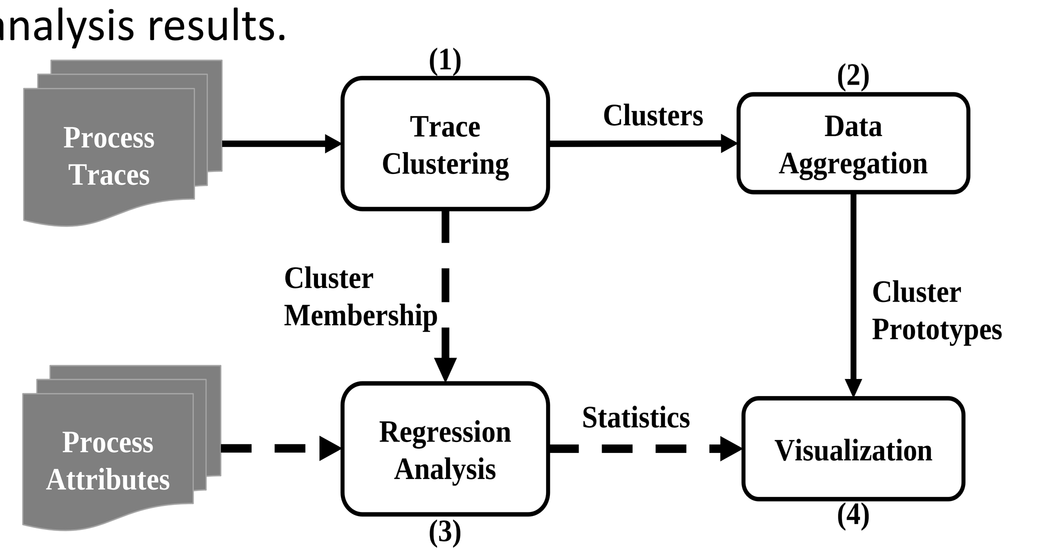
Process sequencing is necessary before more advanced processing. Activities coded in a process log usually have start and end timestamps (some logs may not include end time) for each activity. Idle time may exist between activities, and some activities may be executed concurrently. In process mining, process traces are usually sequenced by ascending order of the start time of activities.



Method

The core methods in VIT-PLA can be summarized as:

1) Clustering of process traces (workflow data)
2) Aggregation of process traces and selection of cluster prototype (representative of the other traces in the cluster)
3) Regression analysis to explore underlying knowledge
4) Interactive visualization of process traces and statistical analysis results.



Method

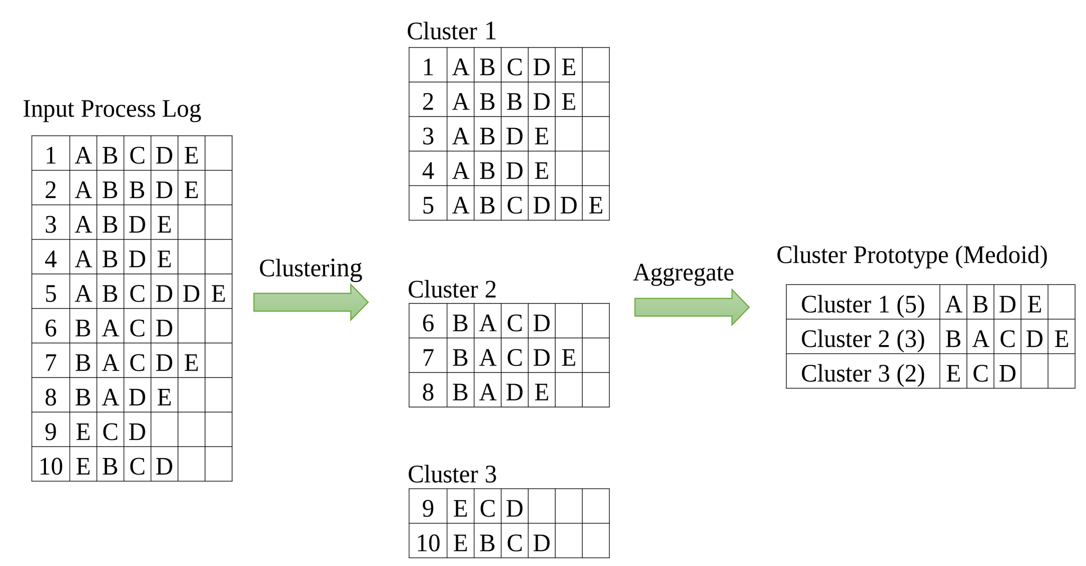


Figure 1. An example showing data clustering and aggregation. The cluster prototype used here is cluster medoid.

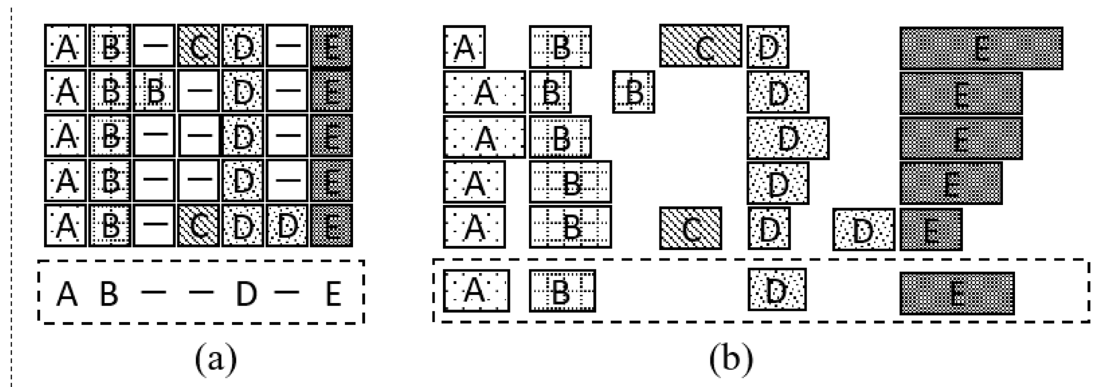


Figure 2. An example of two types of trace alignment: (a) Context-Aware and (b) Duration-Aware. The sequences at the bottom of (a) and (b) are consensus sequences derived from the data. A gap symbol "-" or white space is inserted if a match cannot be found. The five process traces shown here are from Cluster 1 in **Figure 1**.

Multinomial Logistic Regression (one-vs.-one):

$$\ln\left(\frac{P(C=i)}{P(C=J)}\right) = \ln\left(\frac{P(C=i)}{1-\sum_{j=1}^{J-1} P(C=j)}\right) = \beta_{i0} + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdots + \beta_{iK}x_{iK}, \qquad i = 1, \dots, K-1 \qquad (1)$$

Binomial Logistic Regression (one-vs.-rest):

$$\ln\left(\frac{P(C=i)}{P(C\neq i)}\right) = \ln\left(\frac{P(C=i)}{1-P(C=i)}\right) = \beta_{i0} + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \cdots + \beta_{iK}x_{iK}, \qquad i = 1, \dots, K \qquad (2)$$

Figure 3. VIT-PLA GUI design

Preliminary Case Study I

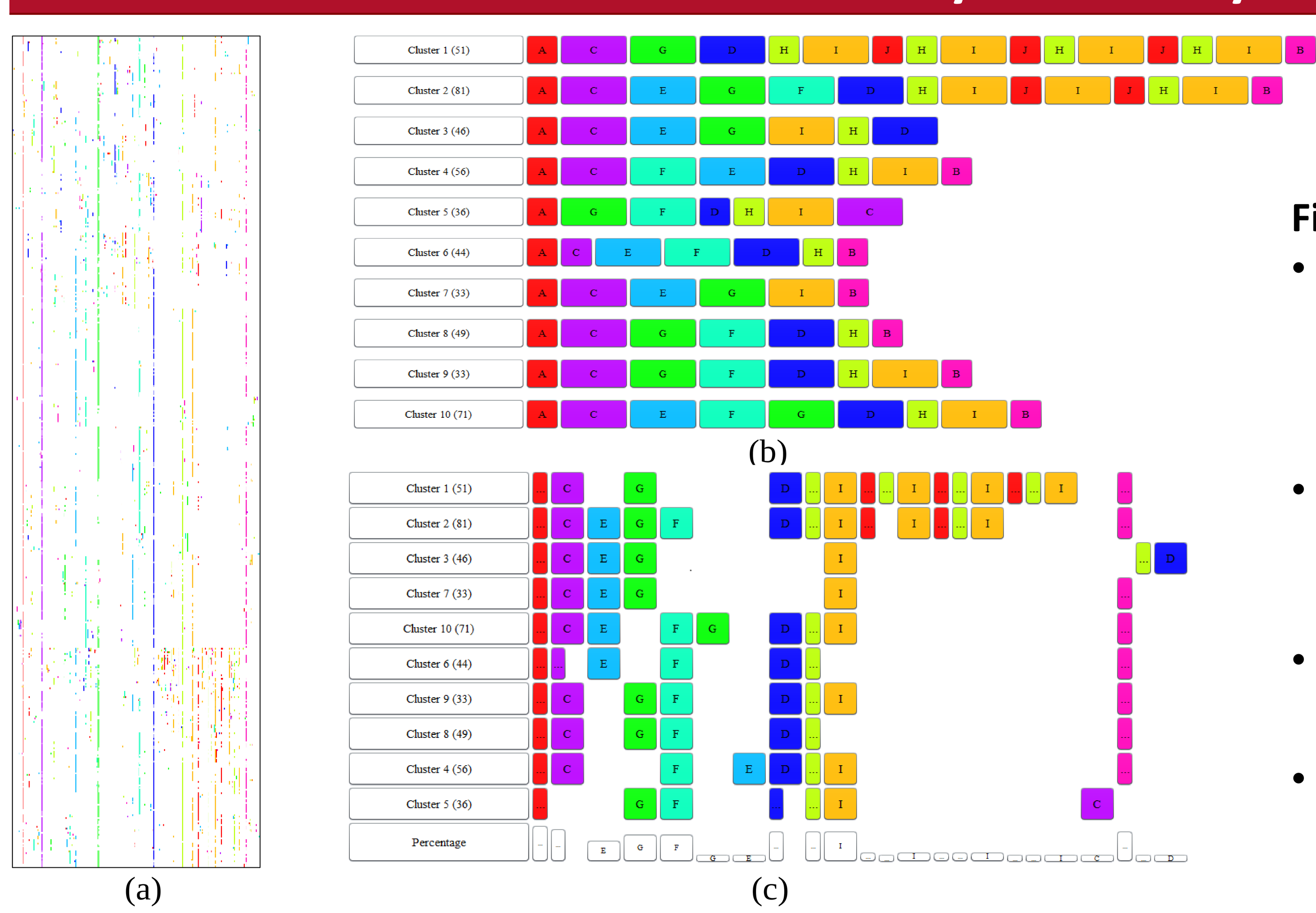


Figure 4. Visualization of artificially generated dataset. (a) Alignment view of all 500 process traces; (b) Simplified visualization of 500 process traces using 10 cluster prototypes; (c) Alignment view of 10 cluster prototypes.

Findings:

- Sequential order of consensus tasks (tasks that occur more than or equal to 50% in the column) is "ACEGFDHIB";
- Pattern "HIJ" is repeated in two of the ten clusters (cluster 1 and cluster 2);
- Activity C is performed late in one cluster (cluster 5);
- Activity D is performed late in one cluster (cluster 3) and omitted in another (cluster 7)

Case Study II

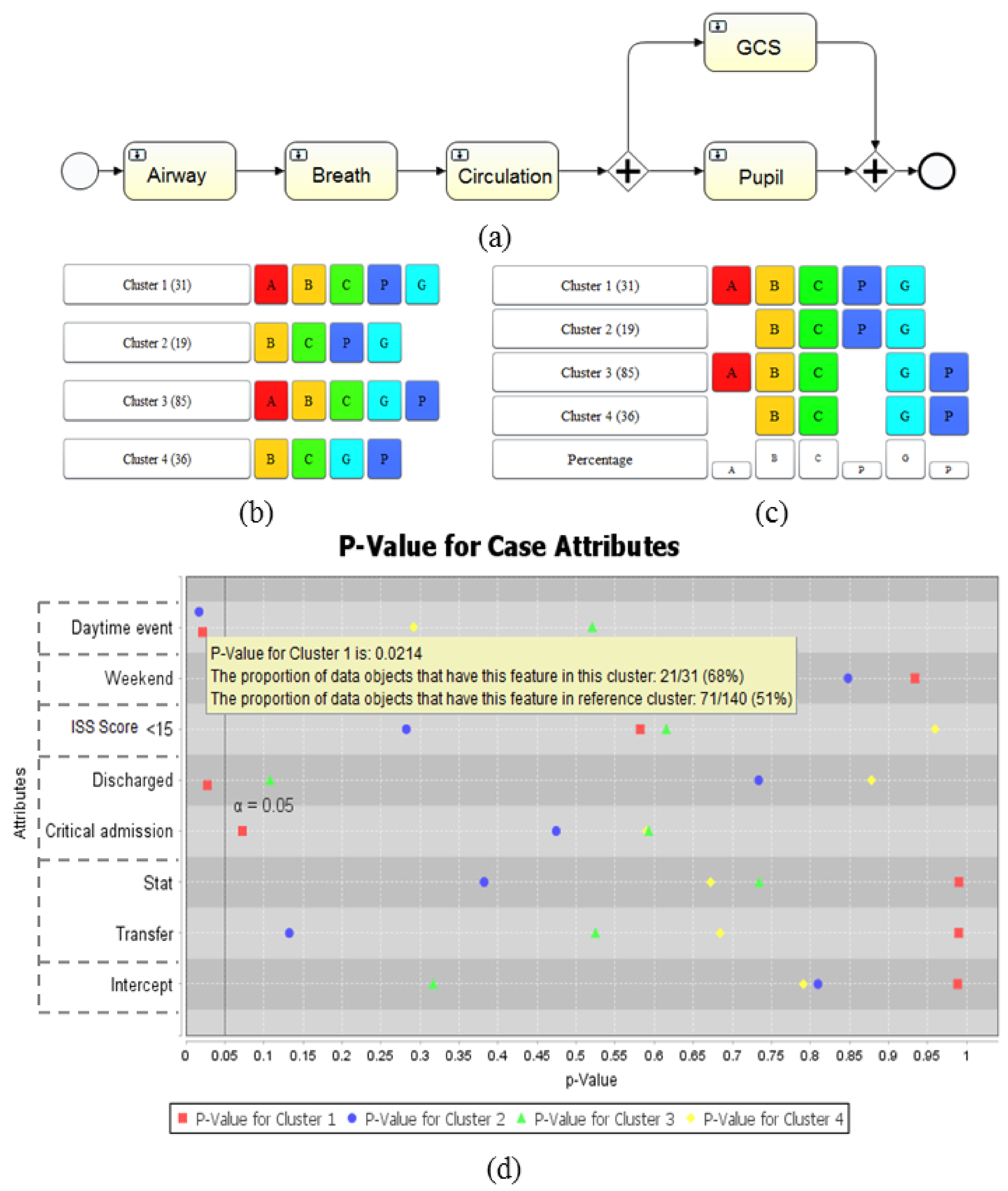


Figure 5. (a) Workflow model (drawn based on BPMN) given by domain expert describing the initial evaluation of trauma, (b) Simplified visualization of 171 traces using four cluster prototypes, (c) Alignment view of four cluster prototypes (d) *p*-value for binomial logistic regression coefficients

Findings: We find that one cluster (cluster 1) whose cluster prototype follows the model occurs more often during the day and another cluster (cluster 2) whose cluster prototype deviates from the model occurs more often at night. This association finding supports previous work showing decreased compliance with trauma protocols (ATLS) at night. (Carter, Elizabeth A., et al. 2013)

Future Direction

In upcoming studies, we seek for:

- In addition to hierarchical clustering, we will evaluate other clustering algorithms (e.g. KNN, k-means, HMM-based clustering).
- Instead of deciding the cluster number manually, we plan on building a function that suggests cluster number based on some cluster metric.
- Apply our approach on more real world datasets.

Acknowledgement

This research is supported by National Institutes of Health under grant number 1R01LM011834-01A1.

RUTGERS

Children's National

Drexel UNIVERSITY

Software: VIT-PLA is still under development, we have not published it yet. Please contact Sen Yang (sy358@scarletmail.rutgers.edu) if you have workflow data that you would like to analyze with VIT-PLA