Thesis Proposals – Trace Clustering with *k-traceoids*

Umut Nefta Kanilmaz
Seidl Lab @ Data Base Systems and Data Mining AI Group

07.12.2025

Overview

The Data Base Systems, Data Mining, and AI Group is offering thesis projects around a recently developed trace clustering framework, *k-traceoids*[1], to motivated Bachelor and Master students. The proposed topic requires implementation of algorithm extensions in the Python programming language. The goal is to perform an extensive experimental analysis and evaluation that contributes to the understanding of *k-traceoid's* properties.

If you are interested, please reach out to kanilmaz@dbs.ifi.lmu.de providing a short description of your background, your prior programming experience, and a transcript of records of current grades.

Introduction

Traces represent ordered sequences of events, where each event belongs to an unique instance and records an activity execution with a given timestamp. These traces capture real-world processes and are often derived from sources like event logs, which stem from process execution data. While traces can provide valuable insights, they can become highly complex as the number of distinct activities and transitions between activities increases (see Figure 1 on the left). This complexity poses a challenge for analyzing trace data effectively.

To tackle this, trace clustering techniques aim to identify meaningful patterns within complex trace data by grouping similar traces together. This allows for the discovery of underlying structures that might otherwise be obscured by the data's complexity.

One such technique we have developed is *k-traceoids*. Inspired by the k-means algorithm, *k-traceoids* has distinct advantages by not relying on encoding traces into a vector space: It preserves the original structure of the trace and

[1] `https://github.com/NeroCorleone/k-traceoids`

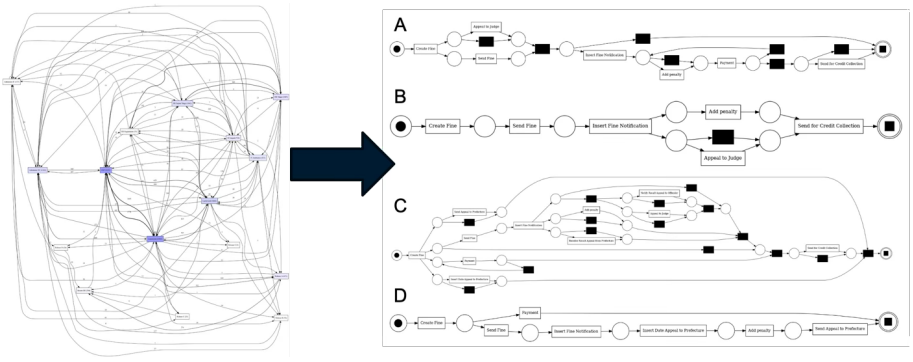

Figure 1: Representation of a real world events log (left) and a visualization of the resulting clusters (right).

keeps the representation of the underlying process. For more details, please refer to the relevant publication[2].

All of the outlined thesis projects contain the following tasks:

- Review of relevant literature
- Setup of code and Python development environment
- Algorithm extension implementations (details below)
- Experimental analysis (details below)
- Documentation and thesis writing

BA Thesis Proposal: Extend Available Model Represention in *k-traceoids*

In the *k-traceoids* algorithm, model representations of the traces within each partition play a central role. Currently, *k-traceoids* supports three different model discovery algorithms: Inductive Miner Infrequent, Heuristic Miner, and Directly Follows Graph which each, in turn, yield a model-based representation of the partitions. However, we are particularly interested in evaluating an additional model based on Integer Linear Programming[3] to determine its effectiveness in the clustering process.

The main goal of this thesis is to assess how well this model performs compared to the existing models, specifically in terms of clustering quality and runtime efficiency. The following tasks are involved in this thesis:

- **Implementation:** First, the framework needs to be extended by the ILP Miner as model centroid representation. The goal is to experimentally

[2] `https://ml4pm.di.unimi.it/preproceedings/ICPM_2025_paper_213.pdf`
[3] `https://link.springer.com/chapter/10.1007/978-3-540-68746-7_24`

evaluate and compare this new model against two other exiting models, namely

a) Inductive Miner Infrequent (existing)

b) Directly Follows Graph (existing)

c) ILP Miner (needs implementation)

- **Evaluation:** The evaluation will be performed on three different datasets to assess the performance and behavior of the new model.

- **Expected Outcome:**

1. A quantitative comparison of cluster quality metrics such as fitness, the number of iterations required for convergence, execution times, etc. for each model and each dataset.
2. A qualitative comparison of the clustering outcomes to understand of how the newly implemented model influences the overall clustering results.