Thesis Proposals – DBSCAN Clustering for Traces

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 a Bachelor thesis project related to the topic of trace clustering. Motivated by a recently developed framework, *k-traceoids*[1], we are now interested to compare against density-based clustering approaches such as DBSCAN. The goal of this thesis is therefore to contribute an extensive experimental analysis and evaluation and comparison of clustering outcomes of trace-adjusted DBSCAN against k-traceoids. The proposed topic requires a small amount of implementation and experimental evaluation preferably with the Python programming language.

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.

We recently explored ways to cluster traces by a novel algorithm inspired by k-means, namely *k-traceoids*[2] that works on traces and preserves their strucutre

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

[2] `https://ml4pm.di.unimi.it/preproceedings/ICPM_2025_paper_213.pdf`

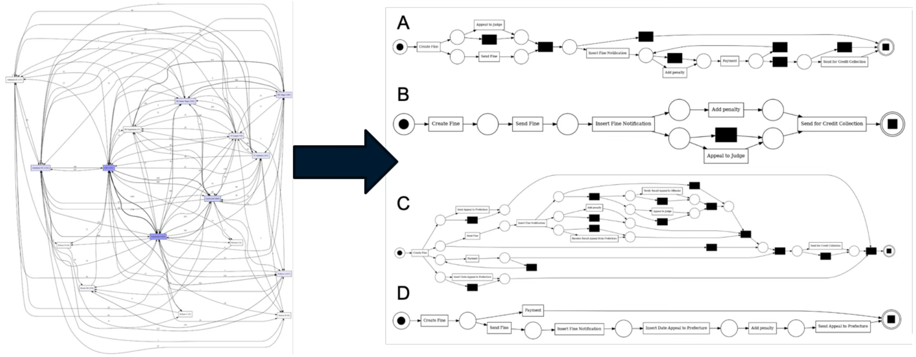

Figure 1: Representation of a real world events log (left) and a visualization of a resulting clusters (right).

(no vectorial encoding). This work motivates us to explore if and how traces can be clustered with density based clustering algorithms. The goal of this thesis project is therefore to identify how the DBSCAN algorithm procedure can be extended to work on traces.

The thesis project contains the following tasks:

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

How can DBSCAN be adapted for trace clustering?

DBSCAN groups points based on density by identifying core points that have at least minPts neighbors within a specified radius ϵ. Clusters are expanded through density-reachable points and points that are not reachable from any core point are labeled as noise. In this way, DBSCAN can identify clusters without a predefined number of clusters. For the means of this thesis, the key adoption lies in the distance function and the definition of the radius.

The following steps are a rough outline on how to approach this thesis:

- **Choice of distance function and other hyperparameters:** The goal is to review suitable distance functions that can work on traces along with sensible hyperparameter choices that will be considered with the experimental evaluation.

- **Implementation:** This task involves setting up the coding environment and implementing the experimental evaluation.
- **Evaluation:** The adjusted DBSCAN will be evaluated on at least three different datasets and compared against three different competitors (including k-traceoids), to assess the performance and behavior.
- **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 initialization strategy for three datasets and X competitors.
  2. A qualitative assessment of the clustering results, providing insights into how well interpretable the clustering results with adjusted DBSCAN are.