Trace Clustering User Guide

General Information

Trace Clustering is a basic application used to discover clusters of traces (sequences of events) with similar characteristics in an event log. The software is provided as a locally hosted standalone application accessible through most standard web browsers, and is intended for all users seeking to find clusters based on a user-provided sample of the cluster.

The application detects traces in an event log based on a user-provided selection of sample traces. Based on the sample, the software discovers *frequent sequence patterns* (FSPs), which are patterns of sequences that appear often in the traces provided, and presents these FSPs to the user. The user shall then confirm certain thresholds (detailed later in this guide), and the software will scan through the entire event log and return traces that match those criteria. The user may then download the resulting cluster as a .xes or .csv file.

Download and Installation

System Requirements

A Windows, Mac or Linux-based Operating System with minimal hardware specifications is required to run this software, along with:
- a Docker installation on your Operating System
- a (highspeed) Internet connection, to download the software
- a Javascript-enabled web browser
- at least 2GB of free hard disk space

Downloading the Docker Image of the Software

1. Ensure that Docker is installed on your computer and a highspeed internet connection is available.
2. Open your computer's command line application (Terminal on macOS/Linux, Command Prompt on Windows), and type `docker pull tongtghs/clustering-app:latest` into the command line, and press "Enter" on your keyboard to run the command. Your computer will start downloading the Docker Image of the Software from Docker Hub. This process may take several minutes, depending on the speed of your internet connection.
3. After the docker image has been successfully download to your computer, the command line returns to its default state. No further installation steps are required.

Using the Software

To run the software, ensure that your web browser (Google Chrome/Mozilla Firefox are suggested) is updated and has Javascript and cookies enabled. Then, the software may be started by one of two ways:

Method 1: Starting the Application from Docker Desktop

1. Open Docker Desktop
2. Select the container labeled "tclustering-app" from the list of containers.
3. On the top right of the screen, press the RUN button.
4. Copy the URL (`http://0.0.0.0:5001/`) and paste it in your browser. Press "Enter" on your keyboard. The application should now be displayed as a web page.

Method 2: Starting the Application from Command Line

1. In your computer's command line application (Terminal on macOS/Linux, Command Prompt on Windows), type `docker run -p 5001:5001 tongtghs/clustering-app` and press "Enter" on your keyboard to execute the command.
2. Copy the URL (`http://0.0.0.0:5001/`) and paste it in your browser. Press "Enter" on your keyboard. The application should now be displayed as a web page.
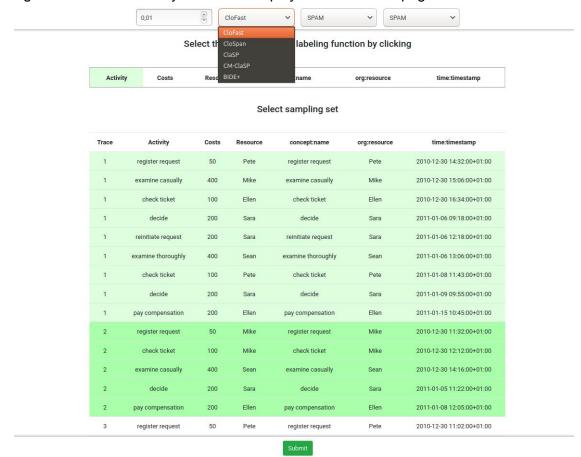
Upload .XES or .CSV File

1. On the homepage of the application, click on "Choose XES or CSV file". Select your log file, in either .XES or .CSV format, and click "Open"
2. Press "Upload File".

Trace Clustering

Choose different XES file

Upload file

Select Labelling Attributes and Sample Cluster

Your Event Log will now be displayed on the browser as a scrollable table. Now, the sample traces must be selected from the event log by the user and a Support value specified:

1. On the Event Log table, click on traces that belong to the sample cluster. Selected traces will be highlighted in light green.
2. Select of the attributes (column names of the Event Log) that are to be used to identify each row. Selected attributes will be highlighted in light green.

3. In the top-left field, enter a numerical value between 0 and 1 for the support of the Algorithm.
   a. Support: the number of traces in a set of traces that contain a sequence pattern.
4. Press the "Submit" button at the top of the page to confirm your sampling cluster. The algorithm will now identify FSPs and display them on a new page.

Viewing and understanding Frequent Sequence Patterns
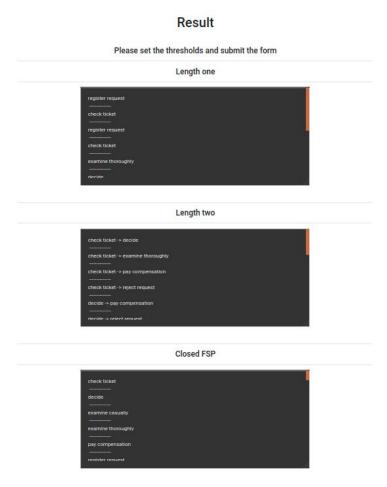
Explanation of Frequent Sequence Patterns
After pressing the "Submit" button in the previous step, you'll automatically be redirected to a "Results" page that displays the identified *frequent sequence patterns* (FSPs). FSPs are simply *sequences of activities* that occur "frequently" in traces from the provided sampling cluster. The Algorithm identifies three categories of FSPs:
- Length one: these are sequence patterns in your sampling cluster that consist of a single activity.
- Length two: sequence patterns in your sampling cluster that consist of two consecutive activities.

- Closed Sequence Pattern: sequence patterns of any length that have a higher support than any other sequence patterns that contain it. In other words, a closed sequence pattern is just a sequence pattern that occurs most often in the provided sampling cluster.

As these FSPs are simply *sequences of activities* that appear often in your sampling cluster, they characterize the traces in the sampling cluster. Other yet-to-be identified traces in the event log will be evaluated against these FSPs: a trace will be included in the final Clustering, if it contains enough of the FSPs (in other words, traces that contain the same FSPs as those in the sampling cluster are considered *similar* to the sampling cluster, and therefore included in the Clustering).
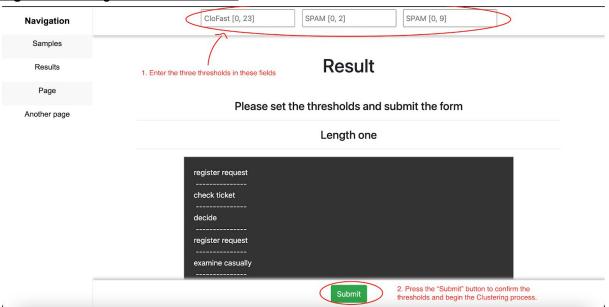
Therefore, based on the displayed FSPs on the result page, the user must now specify three appropriate *thresholds* for the algorithms. A *threshold* is simply the minimum number of FSPs that a trace must contain in order to be included in the cluster. The higher the threshold, the more "similar" a trace in the event log has to be to the sampling cluster, in order to be included in the final Clustering.

Result

Please set the thresholds and submit the form

Length one

register request
----
check ticket
----
register request
----
check ticket
----
examine thoroughly
----
decide

Length two

check ticket -> decide
----
check ticket -> examine thoroughly
----
check ticket -> pay compensation
----
check ticket -> reject request
----
decide -> pay compensation
----
decide -> reject request

Closed FSP

check ticket
----
decide
----
examine casually
----
examine thoroughly
----
pay compensation
----
register request

Setting Cluster Thresholds

1. Enter the three thresholds, as you determine appropriate, in the three fields on top of the page.
2. Press "Submit" to confirm the thresholds. The application will now scan through the rest of the Event Log and identify any trace that meets the thresholds.

Figure 3: Setting Cluster Thresholds

Saving Results

The final Clustering will now be displayed on a page. Press "Download Results" to download the Clustering as a .XES or .CSV file for later viewing on your computer.