Interactive Process Clustering with t-SNE

Steffen Schuhmann[1,2], Jana-Rebecca Rehse[1,3], Sebastian Baumann[1,2], and Peter Fettke[1,2]

[1] German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
{firstname.lastname}@dfki.de
[2] Saarland University, Saarbrücken, Germany
[3] University of Mannheim, Mannheim, Germany
rehse@uni-mannheim.de

Abstract. Process trace clustering is a well-studied and powerful technique to support the discovery of high-quality process models. It splits an event log into more cohesive sublogs, such that the discovered process models are easier to read and to understand. However, existing clustering approaches typically optimize measures like fitness or precision instead of focusing on the model understandability and utility, as assessed by a process analyst. In addition, they offer no opportunity to influence or adapt the clustering result according to the analyst's use case or preferences. In this paper, we propose an interactive tool to trace clustering based on the t-SNE algorithm. Traces are represented in a two-dimensional graph, where they can be selected interactively for process discovery. We also offer the user some guidance with a predefined selection of possible clusters. Using this system, a process analyst is able to find a representative set of process models for each event log without any knowledge in programming and a basic understanding of the used discovery techniques.

Keywords: Trace Clustering · Process Discovery · Process Analytics · Interactive Data Analytics

1 Introduction

The main goal of process discovery is to visualize a real-life business process, as recorded in an event log, in a human-readable way [7]. In reality, however, discovery approaches often produce spaghetti models, i.e., highly complex models that are difficult to read and to understand [2]. Spaghetti models originate in overly complex process logs. For example, if the process contains multiple variants for handling different types of business objects, all of those variants need to be included in the discovered model [7]. In this case, it makes sense to split the event log into multiple logs and discover a separate model for each of them [1]. The challenge lies in determining the best way to split the log, such that we end up with a minimal number of maximally useful models. For this purpose, process trace clustering is a well-known and effective technique, which has been extensively studied and applied in many contributions, e.g., [2, 3, 6].

However, those existing clustering approaches typically optimize measures like fitness or precision, whereby model understandability and utility are considered after generating the process models of the clusters. In addition, they offer the analyst no opportunity to influence or adapt the clustering result according to the concrete use case or preferences and they are often not integrated with process discovery.

Therefore, we designed a novel interactive process clustering (IPC) tool based on the t-SNE algorithm [8]. This algorithm is well suited for embedding high-dimensional data, such as a trace similarity matrix, into a lower dimensional space. The embedding of such a similarity matrix can then be visualized in a two-dimensional graph, where traces with a high similarity are placed closer together.

We integrated this clustering technique into an interactive web-based tool, where the analyst can influence the clustering parameters, select clusters of traces, discover models for those clusters, and compare their similarity. Moreover, we included process discovery algorithms to compute a set of process models that appropriately represent the event log. Compared to existing clustering tools, IPC is both interactive and visual, giving the process analyst a useful guidance tool with a high degree of freedom. The visual representation of the two-dimensional embedding leads to a better understandability of the coherence in the event log. Also, the free selection provides the user with the ability to select groups based on the utility of the concrete use case.

2 Main Characteristics and Innovation

The objective of the IPC tool is to provide process analysts with an easily understandable and interactive visualization of trace similarities, to find an appropriate set of process models to represent the log at hand. As outlined in Fig. 1, the underlying approach consists of four major steps, which are either backend computations or frontend interactions between the tool and the user. In the first step, we compute the pairwise similarity between all traces in the log. The resulting similarity matrix is used as the basis for the t-SNE embedding in the second step. This embedding is then visualized in a two-dimensional graph, which the process analyst can use to gain insights into the log and either manually select clusters for which to discover a process model or have the tool suggest clusters automatically. The set of discovered models is evaluated by comparing their similarity, following the idea that the less similar two models are, the more sense it makes to keep them as separate models.

To be easily accessible without a complex installation process, the IPC prototype[4] is implemented using web technologies and therefore usable with any modern internet browser. The user interface was designed to support process analysts by providing them the options needed for the cluster analysis without cluttering the UI with too many features or complex options. It is split into two

[4] http://ipc.sschuhmann.de/

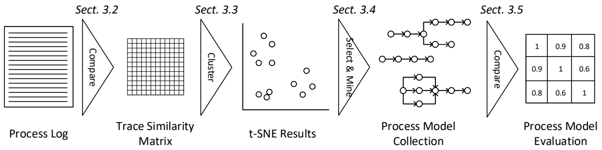


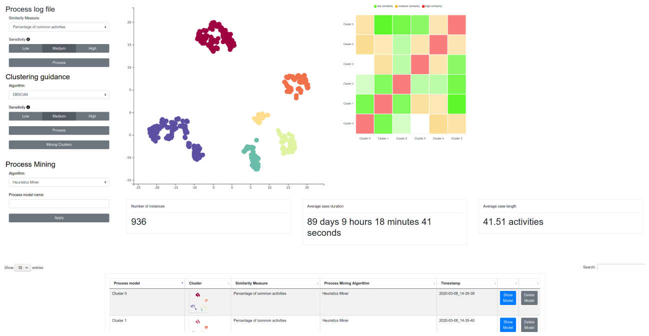
Fig. 1. Main Idea for our Interactive Process Clustering (IPC) Approach

main screens. Users first see the process log upload prompt. It contains a file picker to upload an event log in the XES[5] format. The second screen, shown in Fig. 2, contains all elements used for the clustering process and can be divided into five main groups. On the left, the parametrization section contains all buttons for parametrizing and starting the t-SNE visualization, the clustering guidance, and the process discovery for a selected cluster. Next to it, there is a two-dimensional scatterplot, where the t-SNE results are displayed. It gives the user the option to select a subset of processes as a cluster by dragging a bounding box around it. After at least one process model was discovered for a selected cluster, the similarity matrix based on percentage of common activities is displayed on the right side. It shows the similarity between all generated process models in a color-coded matrix, where green elements highlight a low similarity and red elements indicate a high similarity between the process models. This color scheme originates in the goal to find process models that are as distinct as possible. Below the plot and the similarity matrix, there are three boxes containing descriptive data about the selected process instances, namely number of instances, average case duration, and average case length. Initially, these boxes display information about all contained traces.

The lowest section of the user interface shows the process model table. This table contains the generated models along with their metadata. These include the name, which was used to generate the process model, an image of the plot highlighting the selected cluster used to generate the process model, the similarity metric used to generate the embedding, and a timestamp indicating the time of the model generation. The table also contains two interaction buttons, “Show Model” and “Delete”. This former displays the generated process model in an overlay, the latter removes the process model from the table.

IPC is an easy-to-use tool for discovering process models from event log clusters by interactively selecting the clusters on a two-dimensional projection. This projection changes according to the chosen similarity metric and therefore represents different aspects of the event log, such as the similarity of the traces' structural composition. There are few other contributions in the process mining field that emphasize the visualization of trace clustering results, using, e.g., t-SNE. Schirmer et al. use it as a tool for event log pre-processing [5]. Their approach is similar to ours in terms of visualization and similarity measures, but

[5] http://www.xes-standard.org/



Process log file

Similarity Measure: Percentage of common activities

Sensitivity: Medium

Clustering guidance

Algorithm: DBSCAN

Sensitivity: Medium

Process Mining

Algorithm: Heuristics Miner

Process model name:

Number of instances: 936

Average case duration: 89 days 9 hours 18 minutes 41 seconds

Average case length: 41.51 activities

Show 10 entries

Search:

Process model	Cluster	Similarity Measure	Process Mining Algorithm	Timestamp		
Cluster 0		Percentage of common activities	Heuristics Miner	2020-03-08_14-35-39	Show Model	Delete Model
Cluster 1		Percentage of common activities	Heuristics Miner	2020-03-08_14-35-40	Show Model	Delete Model

Fig. 2. The IPC user interface in the clustering process

it is not integrated with process discovery, uses pre-labeled data, and focuses more on finding outliers as a preprocessing step than on interactive process discovery. Different from our approach, it also does not implement a caching strategy, which may lead to computing times of several hours to days, depending on the size of the process log.

3 Tool Maturity

The tool was implemented as a demonstration unit to evaluate the usability and effectiveness of the proposed clustering approach in a user experience study. To ensure the quality of the software and allow the participants to focus on utility and usability, the implementation is build upon well-known frameworks like Flask[6], scikit-learn[7] and PM4Py[8] in the backend and D3.js[9] and jQuery[10] in the frontend. A video providing a brief overview of the work with the evaluation dataset can be found online[11].

The goal of our evaluation was to assess the utility of the IPC approach and the usability of the IPC tool. For this purpose, the 16 participants were given a short introduction to the tool and provided with a publicly available real-life event log. Then, they were asked to find a number of clusters, which they found appropriate for the given log, i.e., which adequately represented the log without producing too complex process models. Afterwards, they were asked to assess the tool using the User Experience Questionnaire [4]. Users in general appreciated the tool, ranking it with a score around 1.8 in attractiveness, efficiency, and stimulation and only a slightly lower score (1.66) for novelty. However, evaluation scores were lower (around 1.2) for perspicuity and dependability.

[6] https://flask.palletsprojects.com/en/1.1.x/
[7] https://scikit-learn.org/stable/index.html
[8] https://pm4py.fit.fraunhofer.de/
[9] https://d3js.org/
[10] https://jquery.com/
[11] https://cloud.dfki.de/owncloud/index.php/s/wb234DfbLAsKmBG

Since some of the operations, like calculating the similarity matrix and the t-SNE algorithm, are computationally complex, we implemented caching system to reduce the number of these operations. This enables the users to run multiple analysis on the same data in a reasonable time. Therefore, we store the calculation results for the similarity matrix and t-SNE calculation on the server. These cached results are accessed by using a salted hash of the original event log. Since the current implementation does not feature a user management, the cached results are accessible for all users with access to the particular dataset. This way, all users benefit from the cached results after the initial calculation.

4 Conclusions and future work

This paper presents our tool for Interactive Process Clustering using t-SNE and manual as well as automatic cluster selection. The tool was developed to validate the usability of t-SNE in business process analysis and its relevance to trace clustering. The currently implemented similarity metrics focus on the structural composition of the process instances. In future work, we will extend those metrics to enable the user to focus on other aspects of the process traces, such as resources or other metadata. We also will include more advanced process discovery algorithm in a later release.

References

1. Bose, R.P.J.C., van der Aalst, W.: Context aware trace clustering: Towards improving process mining results. In: Proceedings of the 2009 SIAM International Conference on Data Mining. pp. 401–412 (2009)
2. De Medeiros, A.K.A., Guzzo, A., Greco, G., van der Aalst, W., Weijters, T., Van Dongen, B., Saccà, D.: Process mining based on clustering: A quest for precision. In: Business Process Management Workshops. pp. 17–29. Springer (2007)
3. Evermann, J., Thaler, T., Fettke, P.: Clustering traces using sequence alignment. In: Proceedings of the 11th International Workshop on Business Process Intelligence,. International Workshop on Business Process Intelligence (BPI-15), located at International Conference on Business Process Management, July 31 - August 3, Innsbruck, Austria. Springer Berlin Heidelberg (2015)
4. Laugwitz, B., Held, T., Schrepp, M.: Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (ed.) USAB 2008: HCI and Usability for Education and Work. pp. 63–76. Springer (2008)
5. Schirmer, L., Campagnolo, L., González, S., Rodrigues, A., Schardong, G., França, R., Lana, M., Barbosa, S., Poggi, M., Lopes, H.: Visual support to filtering cases for process discovery. In: Proceedings of the 20th International Conference on Enterprise Information Systems. pp. 38–49. Scitepress (2018)
6. Thaler, T., Ternis, S., Fettke, P., Loos, P.: A comparative analysis of process instance cluster techniques. In: Thomas, O., Teuteberg, F. (eds.) Proceedings of the 12th International Conference on Wirtschaftsinformatik. Universität Osnabrück (2015)
7. van der Aalst, W.: Process Mining: Data Science in Action. Springer, 2nd edn. (2016)
8. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**, 2579–2605 (2008)