Interactive trace clustering

Thais R. Neubauer, Marcelo Fantinato [co-advisor], Sarajane M. Peres [advisor]

[1] School of Arts, Sciences and Humanities
University of São Paulo (USP)
São Paulo – SP – Brazil

{thais.neubauer,m.fantinato,sarajane}@usp.br

Abstract. *Process mining aims to automatically discover, analyze and improve business processes. Trace clustering is a task commonly used to reduce the inherent complexity of processes by identifying patterns. This research focuses on the application of experts knowledge in process mining through interactive clustering, referred to herein as interactive trace clustering. The aim is to improve trace clustering by reducing potential losses arising from arbitrary assumptions on the similarity between the datapoints, what is commonly required in unsupervised scenarios. Initial experiments considered partitioning clustering and three representation schemes for traces. Preliminary results show potential to improve the trace clustering quality by inserting experts knowledge.*

1. Introduction

Process models are essential tools for achieving success in business management [Weske 2007]. However, because of cultural reasons or the lack of adequate human and material resources, organizations usually do not formalize these models, often leading them to be unaware of the actual process carried out in day-to-day operations. Process mining provides organizations with information on what occurs in their business processes by extracting knowledge from the event logs generated at process execution [Aalst 2016]. Process mining benefits from the knowledge of the data mining field, especially from clustering techniques – one of the three most commonly used data mining tasks in process mining [Maita et al. 2017]. The descriptive nature of clustering allows us to discover patterns and their contexts [da Silva et al. 2016]. Clusters identified in an event log often provide insights on a particular aspect of the process and can be applied to reduce complex problems to simpler ones, making it easier to further work on process mining in its different types: discovery, compliance and improvement.

In process mining, clustering is known as *trace clustering* [Song et al. 2008]. A trace comprises a sequence of distinctly ordered events, i.e., one event occurs before or at the same time as another one. An event log is a set of traces and works as the data source for trace clustering [Lu 2018]. Trace clustering solutions still do not meet all expectations as well as the classic clustering task in data mining. The recent *interactive clustering* approach aims to introduce human expertise into the clustering task [Hu et al. 2014], reducing possible harmful effects from technical decisions (e.g., choice of algorithms, data representation and similarity functions) in the clustering quality [Correa et al. 2015].

This master's project aims to apply interactive clustering to use knowledge of business experts and reduce harmful technical decisions in trace clustering, rising a new field of study named herein as *interactive trace clustering*. We address ambiguities arising from the use of similarity functions as the specific harmful effects to be reduced.

2. Problem definition

Even good trace clustering results, obtained from the use of some arbitrary similarity function, fail to support the discovery of process models useful for the business context, exposing a gap between clustering and business objectives. To reduce this gap, experts knowledge can replace potentially harmful unsupervised decisions. Figure 1 illustrates how trace clustering can often aid in reducing complex problems (red circle)[1] in simpler understanding problems (yellow circle). Even trace clustering results appropriate from the data mining perspective (green circle on the right) may not suffice from the process mining perspective. Based on this context, we established the hypothesis discussed in this research: *interactive trace clustering allows high-quality process mining, from both the data and process mining perspectives (green circles on the left).*

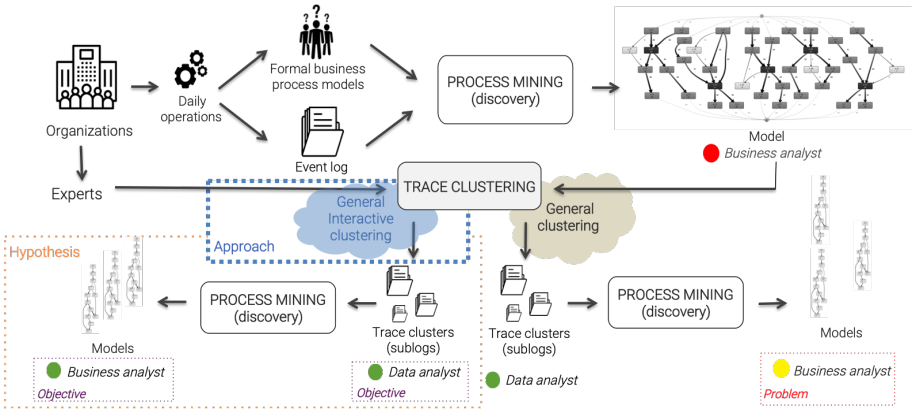

Figure 1. Overview of the *interactive trace clustering* context.

The problem addressed in this project how to reduce the harmful effects of using similarity functions inappropriate to the business context when applying trace clustering. Consider: $X \in \Re^{n \times m}$ the data matrix representing an event log, with n traces and m trace features; $\mathcal{N} = \{\overrightarrow{x_1}, \ldots, \overrightarrow{x_n}\}$ the set of row vectors forming the matrix; k the number of clusters of $\mathcal{N}$ to be found; $\mathcal{K} = \{\mathcal{K}_1, ..., \mathcal{K}_k\}$ the set of clusters of row vectors resulting from the clustering task; W a vector of parameters adjustable through an unsupervised algorithm and $\oplus$ the experts knowledge. Taking the clustering task as the partitioning of the row vectors $\mathcal{N}$ in k parts [Han et al. 2011], *interactive trace clustering* is: $\mathcal{G}_{\mathcal{I}}(X, \oplus) : \mathbb{R}^{n \times m} \times W \rightarrow \mathcal{K}$, where $\mathcal{G}_{\mathcal{I}}$ receives as input the data matrix $X \in \Re^{n \times m}$ and the experts knowledge $\oplus$ and adjusts the vector W to map X to a set of clusters $\mathcal{K}$, such that: $\mathcal{K}_p \neq \emptyset, p \in \{1, \ldots, k\}$; $U_{p=1}^{k} \mathcal{K}_p = X$; and $\mathcal{K}_p \cap \mathcal{K}_q = \emptyset$, $p, q \in \{1, \ldots, k\}$ and $p \neq q$.

3. Research proposal

This work proposes to test *interactive trace clustering* with the algorithm *k-Means++* and two event logs: one with synthetic events and the other with real-world events from IT incident management. Two interactive clustering approaches are being explored: *(i) split/merge*, which uses experts' requests to merge or split clusters [Awasthi et al. 2017], and *(ii) must/cannot-link*, in which the expert determines *must-link* rules for data pairs, when both of them should be assigned to the same cluster, or else *cannot-link* rules [Okabe and Yamada 2009]. Eight experts are collaborating through questionnaires and

[1] Circle colors refer to the satisfaction of business and data analysts: red/low, yellow/medium, green/high

inspection of graphical representations of clustering results. Two experts should synchronously supervise the trace clustering and the others should do it asynchronously.

4. Evaluation

From the data mining perspective, the clustering results should be evaluated through internal (Silhouette [Rousseeuw 1986]) and external (Adjusted Rand [Stanley 2004]) indexes. From a business perspective, indexes adhering to the specific problem of IT incident management should be applied. For process discovery, when applicable, measures of completeness, precision, simplicity and generalization should be applied [Aalst et al. 2012].

5. Acomplished activities

The event logs were pre-processed and mapped to count-based distributional representation schemes (*binary*, *tf* and *tfidf*) [Turney and Pantel 2010]. The logs were clustered using *k-Means++* implemented with Euclidean distance, and the results were analyzed from both the data mining and business perspectives as shown in Figure 2. As expected, the simple use of trace clustering did not show satisfactory results.

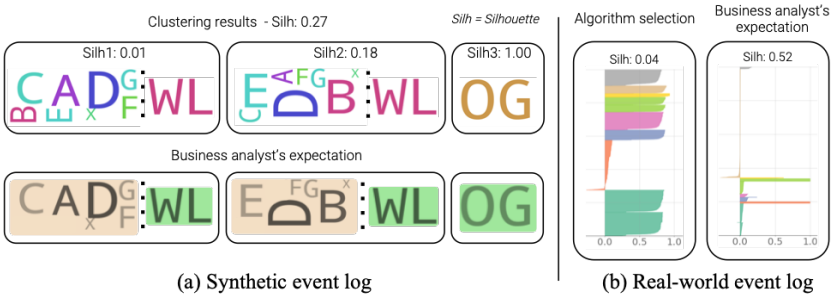

Figure 2. Some of the results visualization regarding the initial experiments.

For the synthetic event log (`https://goo.gl/cGC9U8`), Figure 2(a) shows the clustering results in terms of validation indexes and activities present in the traces belonging to each of the three clusters in contrast to the expectations of the business analyst. The similarity analysis clearly met expectations neither from the data mining nor the business perspectives. As for the real-world event log (`https://goo.gl/EaK96x`), the preliminary analysis considered clusterings performed to support the prediction of incident time resolution as the business goal. The clusterings were performed with trace descriptive features chosen by both an automatic selector and an expert. Silhouette graphs and indexes are shown in figure 2(b). The experts knowledge led to better results from the data mining perspective. From the business perspective, a preliminary evaluation showed a slight advantage for the time resolution predictors built on clusters obtained with features selected by experts.

Next step is to apply interactive trace clustering to improve results in both event logs. For the synthetic event log, *cannot-links* could be imposed on the following pairs of activities: A/B, C/E, A/E and B/C. Possible perspectives that characterize this process are established by traces that include either an activity (e.g., A) or another activity (e.g., B), thus an expert could expect clusters to show such characterizations. For the real world event log, as the business perspective evaluation was performed in terms of prediction error in each cluster, rather than cluster characterization, the *split/merge* approach should be the first attempt. From the evolution of the interaction with experts, visualizations on the clusters characteristics will allow establishing *must/cannot-links*.

6. Final considerations

The results obtained so far show that classic clustering generates results not always adherent to business goals. Therefore, the adoption of an improvement strategy for trace clustering is justified, addressed herein through interactive clustering. In terms of results evaluation, the process mining measures should still be applied for the final validation.

Acknowledgments

Thanks to *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (Capes).

References

Aalst, W. M. P. (2016). *Process Mining: Data Science in Action*. Springer, 2nd edition.

Aalst, W. M. P., Adriansyah, A., and van Dongen, B. F. (2012). Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(2):182–192.

Awasthi, P., Balcan, M., and Voevodski, K. (2017). Local algorithms for interactive clustering. *J. of Machine Learning Research*, 18:1–35.

Correa, G., Marcacini, R., Hruschka, E., and Rezende, S. (2015). Interactive textual feature selection for consensus clustering. *Pattern Recognition Letters*, 52:25–31.

da Silva, L. A., Peres, S. M., and Boscarioli, C. (2016). *Introdução à Mineração de Dados: Com Aplicações em R*. Elsevier, Brasil.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Hu, Y., Milios, E. E., and Blustein, J. (2014). Interactive document clustering with feature supervision through reweighting. *Intelligent Data Analysis*, 18:561–581.

Lu, X. (2018). *Using behavioral context in process mining: exploration, preprocessing and analysis of event data*. PhD thesis, Eindhoven University of Technology.

Maita, A. R. C., Martins, L., Paz, C. R. L., Rafferty, L., Hung, P. C. K., Peres, S. M., and Fantinato, M. (2017). A systematic mapping study of process mining. *Enterprise Information Systems*, 12:1–45.

Okabe, M. and Yamada, S. (2009). Clustering with constrained similarity learning. In *Int. Conf. on Web Intel. and Intel. Agent Tech.*, volume 3, pages 30–33, USA. IEEE.

Rousseeuw, P. J. (1986). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. of Computational and Applied Mathematics*, 20(1):53–65.

Song, M., Gunther, C. W., and Aalst, W. M. P. (2008). Trace clustering in process mining. In *4th Int. Work. on Business Process Intelligence*, pages 109–120. Springer.

Stanley, D. (2004). Properties of the hubert-arable adjusted rand index. *Psychological Methods*, 9(3):386–396.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *J. of Artificial Intelligence Research*, 37(1):141188.

Weske, M. (2007). *Business Process Management: Concepts, Languages, Architectures*. Springer, 2nd edition.