EACH | campus capital USP LESTE
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

Graduate Program in Information Systems
PPgSI

6th PPgSI's Dissertations Workshop 2019

Interactive Trace Clustering

Author:	Thais Rodrigues Neubauer			
Advisor:	Sarajane Marques Peres, Dr.			
Co-advisor:	Marcelo Fantinato, Dr.			
Research lines:	[ ] Systems Development and Management		[x] Systems Intelligence	
Research areas:	[ ] Database [ ] Information technology management	[ ] Software engineering [ ] Human-Computer Interaction	[x] Artificial intelligence [ ] Pattern recognition	[ ] Graphics processing [ ] Optimization
Application areas:	[x] Enterprise environments / Business processes [ ] Economy [ ] Games / Serious games [ ] Health	[ ] Bioinformatics [ ] Education / Distance learning [ ] Linguistics / Natural Language [ ] Other Which? _____	[ ] Biometrics [ ] E-government [ ] Cheminformatics	[ ] Mobile devices [ ] Internet / Social Networks [ ] Robotics [ ] General*
Period in the program (at the workshop date):	[ ] 2nd semester	[ ] 3rd semester	[ ] 4th semester	[x] 5th semester
Qualifying:	[x] Qualifying held in: 26/11/2018		[ ] Plan for qualifying in:	
Defense:	Deadline for deposit: 28/01/2020		Plan for defending in: 30/09/2019	
Publications associated with the master's project:	• NEUBAUER, T. R.; PERES, S. M. ; FANTINATO, M. . Interactive Trace Clustering. In: Workshop de Teses e Dissertações em Sistemas de Informação (WTDSI-SBSI), 2019, Aracaju. Anais do Simpósio Brasileiro de Sistemas de Informação - Workshop de Teses e Dissertações em Sistemas de Informação, 2019.			

The research project summary

Context:

Even good trace clustering results fail to discover useful process models in the business context, exposing a gap between clustering objectives and business objectives. In the process mining context, even if organizations do not have complete knowledge about their business processes models, their professionals (experts) certainly have some relevant knowledge about the processes characteristics. Therefore, the insertion of this knowledge in trace clustering probably brings better results than totally unsupervised assumptions (e.g, pre-determining similarity relations or a type of data distributions). Interactive clustering is a recent approach that focus on inserting the human expert into the clustering task. Applying this interactive approach to trace clustering replaces harmful technical decisions by business expert's decisions. This application defines a new field of study that is being referred to in this project as "interactive trace clustering".

Research problem:

How to minimize the harmful effect caused by similarity functions or data distribution assumptions that are not appropriate to the business context in which the trace clustering is being applied.

Research objective:

Applying interactive clustering to the process mining context to verify how the human's intervention affects the quality of trace clustering results, focusing on the problems caused by ambiguities arising from the use of similarity functions and data distribution choices inappropriate to trace clustering.

Characteristics of the proposed solution:

Interactive trace clustering will be implemented based on partitional clustering algorithms and two interactive clustering approaches: split/merge, which uses experts' requests to merge or split clusters, and 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. Eight experts are collaborating through questionnaires and inspection of graphical representations of clustering results. Two experts should synchronously supervise the trace clustering and the others should do it asynchronously.

Theoretical foundations:

Process models are essential tools for achieving success in business management in organizations. However, due to cultural reasons or lack of adequate human and material resources, it is common for organizations not to formalize these models, frequently making them unaware of the actual process carried out in day-to-day operations. The process mining area provides companies with information about what actually occurs in their processes since the focus of the area is to extract knowledge from the event logs generated during the business process phases. In the process mining context, the clustering task is known as trace clustering. Trace clustering is a popular way of minimizing processes complexity through patterns identification (Maita et al., 2017). Interactive clustering includes a limited amount of supervision in the clustering task to minimize possible harmful influences of technical decisions, such as choice of algorithms, data representation and similarity functions (Correa et al., 2015).

Correlated works:

- Koninck et al., 2017 (An approach for incorporating expert knowledge in trace clustering. In: Advanced Information Systems Engineering): the study applies trace clustering for process discovery and introduces the knowledge of humans (experts) in the solutions for this task.
- Amaral, Fantinato and Peres, 2018 (Attribute selection with filter and wrapper: An application on an incident management process. In: 2018 Federal Conference on Computer Science and Information Systems): it applies automatic attribute selection methods to improve prediction models for incident resolution time. In this work, the authors introduced the event log that will be used in the research presented herein.
- Okabe M., Yamada S., 2010 (Constrained clustering with interactive similarity learning. In: Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems): this study presents an interactive clustering approach that is been used in the research.

Validation

In order to measure the quality of the clustering results from the data mining perspective, internal (Silhouette Index) and external (Adjusted Rand Index) validation indexes will be applied. In order to evaluate results from the business perspective, evaluation indexes adhering to the specific problem under resolution will be applied. Considering the process discovery task, measures of completeness, precision, simplicity and generalization will be applied. The expert's effort will be assessed by calculating the time spent on supervisory tasks and through self-assessments.

Limitations, risks, and threats:

The first limitation identified in this study is the limitation of time that each expert may assign to supervision. Measuring expert effort is a task essentially subjective and relative to each person; therefore, the way it is done in this work may not fully reflect the idea of effort that each person can take for themselves. Moreover, the assessment forms of the intervention on the quality of the resulting groups may not reflect subjective and important aspects such as the alignment of this result with human expectation.

Scientific contribution:

Introduction of interactive clustering to the process mining area, evaluating how expert interventions changes the quality of the obtained results as well as evaluating experts' effort during the supervision tasks.

Technical contribution (if pertinent):
Set of practices formalized in process models for application of interactive clustering to process mining.

The research method			
Genre (choose ONE)	[ ] Theoretical research	[ ] Practical research	[x] Empirical research [ ] Methodological research
Nature (choose ONE)	[ ] Basic research		[x] Applied research
Approach (choose ONE)	[x] Quantitative research	[ ] Qualitative research	[ ] Quali-quanti research
Literature review* (you can choose more than one)	[ ] Narrative review [ ] Descriptive review [x] Scoping review	[ ] Meta-analysis [ ] Qualitative systematic review [ ] Umbrella review	[ ] Theoretical review [ ] Realistic review [ ] Critical review
Main technical procedure (choose ONE)	[x] Experimental research [ ] Bibliographic research [ ] Documental research [ ] *Ex-post-facto* research	[ ] *Survey* [ ] Case study [ ] Participatory research [ ] Research-action	[ ] Ethnographic research [ ] Grounded theory [ ] Design science [ ] Other Which? _____
Data analysis (you can choose more than one)	[x] Descriptive statistics [x] Inferential statistics	[ ] Statistical test [ ] Content analysis	[ ] Discourse analysis [ ] Others: _____

* Definition of types of literature reviews established by Paré, G., Trudel M-C., Jaana M., Kitsiou, S. Synthesizing Information systems knowledge: A typology of literature reviews. In: Information & Management 52, p. 183-199, 2015. DOI: 10.1016/j.im.2014.08.008

Next steps:
Applying split/merge interactive clustering approach into synthetic events dataset and finishing the interactive experiments with the real events log and its experts. In addition, an article (scoping review) will be finalized to be sent to an international journal, an article will be prepared to a conference and the dissertation document also needs to be finished.

Optional: Provide a graphic layout that shows aspects of your research. For example: a flow-chart for building your solution or an infographic for your research proposal. If necessary, use the fourth page.