Generating Artificial Event Logs with Sufficient Discriminatory Power to Compare Process Discovery Techniques

Toon Jouck[1] and Benoît Depaire[1,2]

[1] Hasselt University, Faculty of Business Economics
Agoralaan Bldg D, 3590 Diepenbeek, Belgium
toon.jouck@uhasselt.be; benoit.depaire@uhasselt.be
[2] Research Foundation - Flanders (FWO)
Egmontstraat 5, 1000 Brussels, Belgium

Abstract. Past research revealed issues with artificial event data used for comparative analysis of process mining algorithms. The aim of this research is to design, implement and validate a framework for producing artificial event logs which should increase discriminatory power of artificial event logs when evaluating process discovery techniques.

Key words: Artificial Event Logs; Event Log Simulation; Performance Measurement of Business Processes

1 Research Question

Literature on the comparative analysis of process discovery techniques has revealed some problems with artificial data. The data lacked discriminatory power. We argue that such problems arose due to the absence of a proper framework to generate artificial data. This leads to our main research question: how can we generate artificial event logs with sufficient discriminatory power for a comparative evaluation of process discovery algorithms? To provide an answer to this question several other questions need to be answered:

- What model characteristics can we identify which influence the generated data?
- What is the impact of model language bias on the generated data?
- Which non-model characteristics exist which influence the generated data?
- What is a proper methodology for generating artificial data for comparative analysis?
- Which tools exist for generating artificial data and to what extent are they sufficient?

2 Background

This work focusses on artificial data used for the comparison of different process discovery techniques, more specifically the comparison of control-flow techniques.

In past research on process mining many researchers used artificial data for the development of and the verification of new algorithms (e.g. [1, 2]).

In a recent study De Weerdt et al. compare several process discovery techniques on both artificial and real data [3]. The artificial data used in their experiments was recovered from past research on the development of a process discovery algorithm [2]. Remarkably, the performance of the algorithms did not seem to be significantly different for the artificial data, while real data revealed significant performance differences. These results indicate that the artificial test data used in past research have insufficient discriminatory power.

A lot of process discovery techniques have been developed in the last decade. Since the first algorithms, process discovery has matured remarkably. However, it's still not clear which algorithm will perform best in a certain situation. This has led to an increasing importance of the research on comparing different process discovery techniques [3, 4, 5].

3 Significance

The comparison of process discovery techniques can be based on both artificial and real data. Real data, however, are at a disadvantage when performing such a comparative analysis.

Two disadvantages stem from the nature of algorithm comparison and evaluation. Typically research is performed on a sample of event logs, but conclusions are preferably generalizable to other event logs. To achieve reliable conclusions, statistics require sufficient observations and samples which are representative for the considered population. Real data, however, have limited availability and are typically convenience samples, rather than random samples.

Another disadvantage of real data is concerned with identifying causal relationships between process or event log characteristics on the one hand and algorithms performance on the other. This kind of research requires experimental data and not observational data (real data).

In contrast, these disadvantages are not present when using artificial data in comparative analysis of process discovery techniques if a proper methodology is used to generate the artificial data. Such a methodology should focus on creating artificial data with sufficient discriminatory power to overcome the problems encountered in past research (e.g. [3]). The main contribution of this research will be drawing up and implementing a general methodology for the generation of artificial event logs with sufficient discriminatory power in order to evaluate process mining algorithms.

4 Research design and methods

Firstly, a structured literature review is performed to get insight into generating artificial data and algorithm comparison. The primary sources used to perform

this review are: literature in the domain of process mining and literature from other domains on (generating) synthetic data.

Secondly, the general methodology is built and implemented in a tool to support this new methodology.

Finally the implemented methodology is tested and validated by repeating experiments done in past research on comparing process discovery techniques.

One important limitation of this methodology will be its scope which is limited to generating artificial data for analysing control-flow discovery techniques. Also the reader should be aware that such a general methodology for artificial data will not replace the need for real (test) data. Real logs continue to be necessary for making artificial event logs more realistic and as a final review for process discovery techniques.

5 Research stage

5.1 A Preliminary Framework

The literature review of articles on the evaluation of process discovery techniques based on artificial data (i.a. [1, 2, 3]) reveals that there are only some guidelines or recurring elements for generating artificial logs. However, a sound and general methodology is missing, which decreases the relevance of artificial logs.

To address this issue a preliminary framework is distilled from the literature review which focusses on the crucial aspect of randomization. The methodology can be divided into two stages: the generation of an artificial process model and the generation of event logs from this model. Both stages allow the researcher to define the characteristics of the population and produce a representative sample (see table 1).

The first step is to define a population of process models, from which artificial models are sampled randomly and automatically. In past research this crucial step in generating artificial event logs was never made explicit in a general method or guideline. Mostly processes were drawn manually in an ad-hoc manner without explicitly defining the population they were drawn from. However, it is important that the researcher has insight into the process model population and can influence the properties of that population. Therefore, ranges for the list of controllable properties (see step 1 in table 1) must be set to define the population. Next, values within these ranges are selected randomly and automatically to define a single process model.

The second step concerns the generation of event logs for each process model defined in the previous stage. Again, the researcher must set ranges for several event log properties, from which exact values are sampled randomly to generate event logs. The parameters which can be set are shown in step 2 in table 1.

5.2 Tools for Generating Artificial Event Logs

Different tools already exist which can help to automatically generate artificial event logs. We evaluated two tools considered most appropriate to support the

Table 1. Preliminary Methodology for Generating Artificial Event Logs

Step Methodology	Controllable Properties
1. Model Generation	Number of activity types
	Choice structural patterns
	Choice nested structural patterns
2. Log Generation	Number of generated process instances
	Required completeness
	Noise
	Imbalance of execution properties

preliminary framework: the PLG tool [6] and the BeehiveZ tool [7]. At first sight both tools seem appropriate because both support the two stages of the proposed methodology. However, a more detailed evaluation revealed that both tools do not completely support the proposed methodology and several limitations exist. The results of the evaluation, summarized in table 2, show that the PLG tool supports the preliminary framework the best.

Table 2. Tools for Generating Artificial Event Logs

Properties	PLG	BeehiveZ
Number of activity types	NOK	OK
Choice structural patterns	OK	NOK (indirectly by generator)
Choice nested structural patterns	OK	NOK (indirectly by generator)
Number of generated process instances	OK	NOK
Required completeness	NOK	NOK (only for simple models)
Noise	OK	OK
Imbalance of execution properties	OK	NOK

5.3 A First Step Towards Validation

Although the framework as presented in table 1 is still preliminary, it was used in a first case study to assess if it was a step towards artificial data with more discriminatory power.

For this case study we repeat part of the experiment of De Weerdt et al. [3] in which they evaluated process discovery techniques on both artificial and real event logs. Remarkably, the performance of the algorithms did not seem to be significantly different for the artificial data, while real event logs revealed significant performance differences.

We hypothesize that our methodology can produce artificial data with more discriminatory power. Therefore we repeat part of the experiment of De Weerdt

et al. [3] on artificial data generated with the proposed methodology to see if our results are closer to the results on real data in De Weerdt et al., than their own results on artificial data. If that is true, the case study will provide a first support to our hypothesis.

We applied the preliminary methodology to generate 35 artificial event logs out of two random populations using the PLG tool (with all its limitations). Then four process discovery algorithms were evaluated in two conformance dimensions, fitness and precision, using the method described in [3].

The results in the fitness dimension show that the performance of the tested algorithms reflect better the results for fitness on real data in [3], and thus supports the earlier stated hypothesis. However, the performance differences with respect to fitness in our experiments were of a different order of magnitude than the performance differences based on real data found by De Weerdt et al. [3]. Moreover, the results from our experiments don't show any significant differences in terms of precision in contrast to the results in [3] based on real data.

From these findings can be concluded that the preliminary methodology is only a first step in the direction of increasing the discriminatory power of artificial event logs.

References

1. van der Aalst, W., Weijters, T., Maruster, L.: Workflow mining: discovering process models from event logs. Knowledge and Data Engineering, IEEE Transactions on **16**(9) (September 2004) 1128 – 1142
2. de Medeiros, A.K.A.: Genetic process mining. PhD thesis, Technische Universiteit Eindhoven (2006)
3. De Weerdt, J., De Backer, M., Vanthienen, J., Baesens, B.: A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs. Information Systems **37**(7) (November 2012) 654–676
4. Rozinat, A., de Medeiros, A.A., Gnther, C.W., Weijters, A., van der Aalst, W.M.: Towards an evaluation framework for process mining algorithms. In: BPM Center Report, BPMcenter.org (2007)
5. vanden Broucke, S.K., Delvaux, C., Freitas, J., Rogova, T., Vanthienen, J., Baesens, B.: Uncovering the relationship between event log characteristics and process discovery techniques. In: Business Process Management Workshops, Springer (2014) 41–53
6. Burattin, A., Sperduti, Alessandro, A.: PLG: a process log generator. Technical report
7. Jin, T., Wang, J., Wen, L.: Efficiently querying business process models with BeehiveZ. In: BPM (Demos), Clermont-Ferrand, France (2011)