Merging Computer Log Files for Process Mining: an Artificial Immune System Technique

Jan Claes and Geert Poels

Department of Management Information Systems and Operations Management
Faculty of Economics and Business Administration
Ghent University, Tweekerkenstraat 2, 9000 Ghent, Belgium
{jan.claes,geert.poels}@ugent.be

Abstract. Process mining techniques try to discover and analyse business processes from recorded process data. These data have to be structured in so called *computer log files*. If processes are supported by different computer systems, merging the recorded data into one log file can be challenging. In this paper we present a computational algorithm, based on the Artificial Immune System algorithm, that we developed to automatically merge separate log files into one log file. We also describe our implementation of this technique, a proof of concept application and a real life test case with promising results.

Keywords: Business Process Modelling, Process Mining, Log File Merging

This paper has been published as: Claes, J., Poels, G.: Merging Computer Log Files for Process Mining: an Artificial Immune System Technique. BPM 2011 Workshops, LNBIP. Springer (2011).

1 Introduction

Process mining techniques are used to discover and analyse business processes in a semi-automatic way. Starting from all kinds of recorded process data (called *log files*) process mining tries to automatically discover the structure and properties of the business processes, which can be visualised in business process models.

Three actions have to be taken before process discovery and analysis techniques can be performed: searching for data in the IT support systems, structuring these data (i.e. identifying single process steps (events) and groups of process steps that belong to the same process execution (process instances)), and converting these data to the format required by the process mining tool. If process data are found in different sources, then a fourth action is required: merging the data into one computer log file.

In this paper we present an automated technique for merging already collected, structured and converted process data according to an Artificial Immune System (AIS) algorithm, which is based on the features and behaviour of the vertebrate immune system. By automating this fourth action of the preparation step, we try to broaden the benefits of process mining to an extended part of the overall process mining procedure, because the automation makes the merge step in the preparation phase faster (speed), the use of data from multiple systems is facilitated (completeness) and the way these data are merged is less subjective than when performed manually (correctness).

Our approach is implemented in ProM, a well known academic process mining tool, which implies that for our implementation we assume the different data sets are first separately structured and converted to the ProM file format. Fig. 1 shows the steps for our solution implementation.

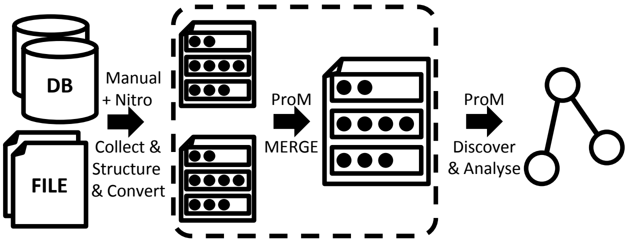

Fig. 1. Merging data of different sources can be performed after structuring and converting to a tool-specific file format. We implemented our merge technique in the ProM analysis tool itself.

2 Technique

The merging of two computer log files consists of two steps: (i) linking together traces of both logs that belong to the same process execution and (ii) merging these traces into one trace to be stored in a new log file. We assume reliable and comparable timestamps are available in the original logs causing the second step to be a simple exercise of chronological ordering of all the events of linked traces into one new trace in the resulting merged log file.

In our opinion, more than one factor can indicate that two traces should be linked We looked for existing techniques that incorporate multiple indicators in their solution procedure and found our inspiration in the Artificial Immune System algorithm. This algorithm uses an affinity score throughout the entire calculation procedure where this score points to the best solution. We used a combined indicators function to derive the affinity score for the algorithm. Each scored solution is not more than a set of linked traces between the two log files. By combining our different assumed indicators, a solution with a high score has a higher chance to be an optimal solution because most combined indicator value points to that solution.

3 Experiment results

We have tested our technique with a simulated example and with a real case example. The benefit of using simulation is that the correct solution (i.e. the process to be discovered) is known end that properties like time difference or noise can be controlled.

The results of both series of tests revealed a nearly perfect merge if both logs used the same trace identifiers. With different trace identifiers in both logs in all cases the correct *number* of links was found, but when traces run partly in parallel, there seems to be too little information left to find the right links. The amount of noise in the logs seems to have little impact on the correctness of the identified links

The full published article can be found at http://processmining.ugent.be/post.php?post=pubbpi2011