An Application of Process Mining by Sequence Alignment Methods to the SAP Invoice Process Example

Jakub Štolfa[1], Svatopluk Štolfa[1], Kateřina Slaninová[1,2], Jan Martinovič[1,2]

[1] Department of Computer Science, FEI, VŠB - Technical University of Ostrava,
17. listopadu 15, 708 33, Ostrava-Poruba, Czech Republic
[2] IT4Innovations, VŠB - Technical University of Ostrava,
17. listopadu 15, 708 33, Ostrava-Poruba, Czech Republic
{jakub.stolfa, svatopluk.stolfa, katerina.slaninova,
jan.martinovic}@vsb.cz

Abstract. Process mining started to be a very useful tool how to check processes in companies. The only thing that has to be done is to have at least some log about the activities performed by the software system. There are many methods that can be used then to analyze this data obtained from the log. Every usage of yet not used or rarely used method opens up new perspectives in process mining and reveals the unknown potential of its application to the practice. In this paper we have applied sequence alignment methods to the real process example and examined what results and benefits could be obtained from such usage. The main purpose of this paper is to adjust methods for sequence alignment to be able to determine similarity between the business processes.

Keywords: Sequence Alignment, Process Mining, SAP

1 Introduction

Generally, information systems support business processes. Enactments of the processes are partly managed by the systems, partly managed by users decisions and activities. It is not easy to understand whether the specific process runs efficiently, because usually various activities are processed in parallel and process definition allows plenty of process enactment variations. Our task was to analyze specific process with request to suggest steps for its simplification, curtailment and enact the process cheaper.

The research described in this paper is based on our previous work that included process reconstruction and path analysis [15]. According to this previous research we were able to adjust the process, recognize the false usage of the process, analyze malfunctions in the reality etc. Other previous research closely related to this paper has been done It was focused on the analysis of the process data that involved usage of the sequence alignment methods [9]. The aim of this paper is focused on adjusting of our approach presented in mentioned

J. Pokorný, K. Richta, V. Snášel (Eds.): Dateso 2014, pp. 81–90, ISBN 978-80-01-05482-6.

previous work to the business process area with the consideration to its specific characteristics, especially to adjust methods for sequence alignment to be able to determine similarity between the business processes. The approach was tested and used in other research areas yet, for example in e-learning area [12], in analysis of behavior of agents during the simulation [14] and in analysis of user behavior on the web [13].

The paper is organized as follows: Section 2 introduces the state of the art; Section 3 describes the process that is analyzed and log obtained from the company, Section 4 depicts the experiment that we have performed, describes the preparation of data that we have obtained, shows the usage of our process mining method and explains obtained results; concluding Section 5 provides a summary and discusses the planned future research.

2 State of the art

Business process definitions are sometimes quite complex and allow many variations. All of these variations are then implemented to supportive systems. If you want to follow some business process in a system, you have many decisions and process is sometimes lost in variations. Modeling and simulations can help you to adjust the process, find weaknesses and bottlenecks during the design phase of the process.

The idea of process mining was introduced by Aalst in 2004 [1, 2]. This area of the research has been developing during the years, lot of methods were introduced to this topic. In 2005, ProM tool was introduces [3]. ProM aggregates methods and approaches in this area of study. There are a lot of papers that describe new ways or improvements of methods, techniques and algorithms used in the process mining, but only several papers are focused on the case studies [8].

In the area of process mining the methods of the sequence alignment were introduced by Esign and Karagoz [4] in 2013. Focus of their work was quantitative approach for performing process diagnostics. The approach uses sequence alignment methods for delta analysis. It is comparison of actually performed process and prescriptive reference model [1]. Our paper provides another usage of the sequence alignment methods. We use these methods for comparison of extracted processes to find similarity in the process executions, i.e. some patterns of the process.

The basic approach to the comparison of two sequences, where the order of elements is important, is The longest common substring method (LCS). This is used in exact matching problems [6]. It is obvious from the name of the method that its main principle is to find the length of the common longest substring. The LCS method respects the order of elements within a sequence. However, the main disadvantage of this method is that it can only find the identical subsequences, which meet the characteristics of substrings.

Unlike substrings, the objects in a subsequence might be intermingled with other objects that are not in the sequence. The longest common subsequence method (LCSS) allows us to find the common subsequence [7]. Contrary to the LCS method, the LCSS method allows (or ignores) these extra elements in the sequence and, therefore, it is immune to slight distortions.

The third selected method was The time-warped longest common subsequence (TWLCS) [5]. This method combines the advantages of the LCSS method with dynamic time warping [10]. Dynamic time warping is used for finding the optimal visualization of elements in two sequences to match them as much as possible. This method is immune to minor distortions and to time non-linearity. It is able to compare sequences, which are for standard metrics, evidently not comparable.

The methods LCS and LCSS used for the comparison of sequences find the longest common subsequence z of compared sequences x and y, where $(z \subseteq x) \wedge (z \subseteq y)$. The relation weight $w_{seq}(x, y)$ between the sequences x and y was counted by Equation 1:

$$w_{seq}(x, y) = \frac{l(z)^2}{l(x)l(y)} \frac{Min(l(x), l(y))^2}{Max(l(x), l(y))^2}, \tag{1}$$

where $l(x)$ and $l(y)$ are lengths of the compared sequences x and y, and $l(z)$ is a length of a subsequence z. Equation 1 takes account of the possible difference between $l(x)$ and $l(y)$. Due to this reason, z is adapted so that $w_{seq}(x, y)$ is strengthened in the case of similar lengths of sequences x and y, and analogically weakened in the case of higher difference of $l(x)$ and $l(y)$. For the methods LCS and LCSS, w_{seq} meets all the similarity conditions: $w_{seq} \geq 0$, $w_{seq}(x, x) = 1$, $w_{seq}(x, x) > w_{seq}(x, y)$ and $w_{seq}(x, y) = w_{seq}(y, x)$.

The output z is only the sequence which characterizes the relation between the sequences x and y for T-WLCS method. Therefore, $w_{seq}(x, y)$ does not meet all the similarity conditions due to its characteristics. Respectively, it is possible that $w_{seq}(x, y) > w_{seq}(x, x)$. Although we know that $w_{seq}(x, y)$ is not a similarity for T-WLCS method, due to a simplification, the 'sequence similarity' will be used as a relation weight $w_{seq}(x, y)$ between the sequences x and y for all the methods of sequence comparison in the following text.

As a complementary method for comparison of sequences we have used common method used in informational retrieval, Leventhtein distance [11].

3 Process Context

The analyzed company runs SAP system in five countries and process approx. 30,000 supplier invoices per year. Examined business process of the invoice verification is implemented in SAP ERP and SAP DMS, user activities are controlled by SAP business workflow. Users participate in the invoice verification workflow in several different roles (creator, accountant completion, approver, and accountant decision and posting). Generally, it is process where the accountant should create the invoice, verify it, send to the approvers and finally when he gets it back he does invoice posting.

We have loaded the log of the process between 1/1/2012 and 6/30/2012, totally we loaded 37,991 records for adjusting. Detailed description of the obtaining log and data preprocessing is described in our previous work [15].

4 An Application of the Sequence Alignment

The main purpose of this paper is to adjust methods for sequence alignment to be able to determine similarity between the business processes. The first step of the experiments was focused on one characteristic of the business processes: the duration of the events. In sequence alignment area, this problem is not common. That is the main reason for adjustment of the methods to this area. We have performed an experiment where we made decision about the categorization of the duration of events, prepared four types of sequences from the different viewpoints, analyzed the prepared sequences and applied the sequence alignment methods to determine the similarities of the sequences. Finally, we have selected the right method for our purposes and analyzed the applicability of our approach on the example. The following subsections describe the particular steps in detail.

4.1 Data Preparation and Categorization - Duration Category

When we look at the histogram of all events that is depictured at Fig. 1, we can see that there are cases that where completed in one day, then cases that last two days etc. We can see interesting distribution of the process duration on this histogram that looks like the waves. The waves are caused by the fact that all events of this process are performed in window of approximately 8 working hours each day. Then there is a 16 hours delay till the next working day window appears.

We needed to categorize the duration of the events. After the consultation with the company, we have decided to set three categories for the event duration. First category is the process that last up to the 168 hours. It means that they were started in the window of one working week and ended the same week or the week after. Next, the second category is the category of processes that last from one week to one month. The last category is the category of all processes that last more than one month. Since there are four events in the process, one process should last up to 36 hours so all four together last up to one week. Similar categorization principle works for other categories too.

We set up aliases for the type of the events in the sequence. It means that Verification event is in the sequence like V, Creation event is C, Approval is A, and Posting is P.

4.2 Types of Event Sequences

According to information from the data log we could analyze information in detail and from different points of view. We have defined four types of points of view, or we can say type of event cases:

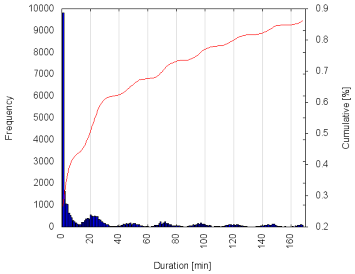

Fig. 1. Histogram of all events

- Type A - Events without Time and Users,
- Type B - Events with Time and without Users,
- Type C - Events without Time and with Users,
- Type D - Events with Time and Users.

Case type A focuses on the topological structure of the process only and does not care about meta-information of the events like time, or duration of the event, or user that performs particular event.

Example of the sequence for the case type A: `0,C;V;A;A;S;`, where sequence is defined by the following structure `CaseID, event1;event2;event3;...;eventn`

Case type B focuses on the topological structure of the process and combines it with the information about time, or duration, of the events. That can bring us another point of view to the process. Thanks to this we can see duration of the whole process, or its events. Tagging of the duration of the events is made by repeating the symbol of the event in the sequence. How much is the symbol repeated depends on the duration category of the particular event. If the event fits in the category three than the symbol is repeated three times, if the event fits in the category two than the symbol is repeated two times and for the category one is symbol repeated one time.

Example of the sequence for the case type B: `0,C;C;V;A;A;S;`, where sequence is defined by the following structure: `CaseID, event1; event1 (repetition according to the duration cat.); event2; ...; eventn`.

Case type C combines topological structure of the process and metainformation about the users. It brings us interesting view to the users involved in particular process.

Example of the sequence for the case type C: `0,C_USER068; V_USER068; A_USER272; S_USER068;`, where sequence is defined by the following structure: `CaseID, event1_originator of the event1; event2_originator of the event2; event3_originator of the event3;...; eventn_originator of the eventn`.

Case type D combines all possible views to the process - topological, time view and users view. It can bring us superb and surprising view to the process that is not possible to see at the first glance.

Example of the sequence for the case type D: `0,C_USER260; C_USER260;V_USER260;A_USER074;A_USER202;S_USER260;`, where sequence is defined by the following structure: `CaseID, event1_originator of the event1(repetition according to the duration category); event2_originator of the event2; event3_originator of the event3;...; eventn_originator of the eventn`.

Thanks to the different case types we can discover processes that take least time, most time to accomplish it, or we can find deviations in the process execution that is not possible to see only in the topological view. We can see if some users take more time to execute event, or if some users communicate more or less, etc.

4.3 Information about sequence types sequence distribution

We have made sequence distribution diagrams for four case types. Some basic interesting information about particular process can be identified here.

Fig. 2 on the left side shows histogram of the most frequented sequences. Histogram nicely shows distribution of the frequency of the particular sequences. It is focused to the case type A. We can see that most frequented sequence in the examined process is sequence `C,V,A,A,S`. Second most frequented is `C,V,A,A,A,A,S`, etc. We can also discover the least frequented sequences that can be identified as some deviations in the process.

Right side of the Fig. 2 shows histogram which is focused to the case type B. There we can see that time (duration of events) influences the amount of different sequences.

Fig. 3 depictures on the left side histogram of case type C and on the right side case type D the sequences with and without time, but with the identification of users (performers of the events). This information also influences the difference between the sequences.

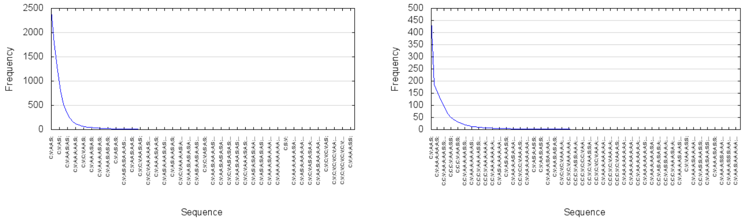

Fig. 2. Left side - histogram Type A, right side - histogram Type B

4.4 Application of sequence alignment methods

We have used four methods to determine the similarity of sequences. Similarity was determined by the equation 1. The main purpose for determining the sequence similarity is that we would like to find similar types of sequences (set

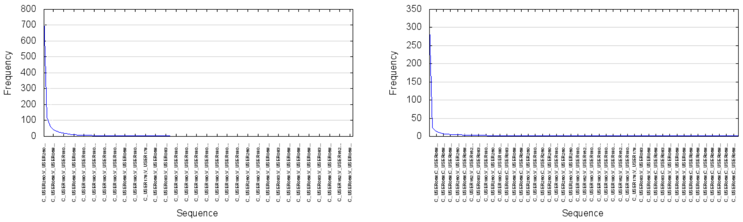

Fig. 3. Left side - histogram Type C, right side - histogram Type D

business processes) as well as the deviations. The four different types of sequences allow us a whole new insight into the performed processes that cannot be done by conventional approaches.

Similarity between sequences was determined by several selected methods. The aim was to find the score, which can show us how similar the event sequences in the case are. The weight distribution in Fig. 4 and Fig. 5 shows us that each method has its specific behavior and due to this the score which determines the similarity between the sequences is different for each method. The advantages and disadvantages describes the following text.

Levenstein distance method does not respect the order of events within the sequences. It only represents amount of necessary steps to change one sequence to another. LCS method respects the events order within the sequences. However, it is suitable, when we can find identical sequences. If the sequences have even a small difference, the similarity weight changes significantly. LCSS method is more tolerant to slight distortions inside the sequences. Of course, it respects the order of events within the sequences. In comparison with LCS, if we add one more event into the sequence, we can find the change of weight similarity, but not as significant then using LCS method. T-WLCS method has as similar behavior as LCSS method. Besides the event order in sequences, T-WLCS method emphasizes the event recurrence within the sequence. However, we must remind that the sequence weight is not similarity for T-WLCS method, it is only a score.

Histograms at Fig. 4 and Fig. 5 show us a distribution of weights for all four case types. These histograms show in fact the applicability of used methods to our example. We can see that the Levensthein and LCS methods are not very suitable for our purposes, because the distribution for these methods does not show much variability, weight distribution is closely concentrated. LCSS and T-WLCS methods seemed to be more promising and we have analyzed the similarities between the sequences in more detail.

Analysis of the result showed that T-WLCS is seems to be more appropriate for our purpose. We have based our decision also on our previous researches that showed suitability of these methods [29, 30].

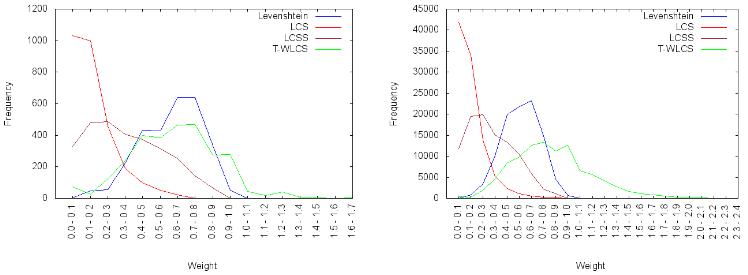

Fig. 4. Histogram - comparison of the methods - Type A(left side), Type B(right side)

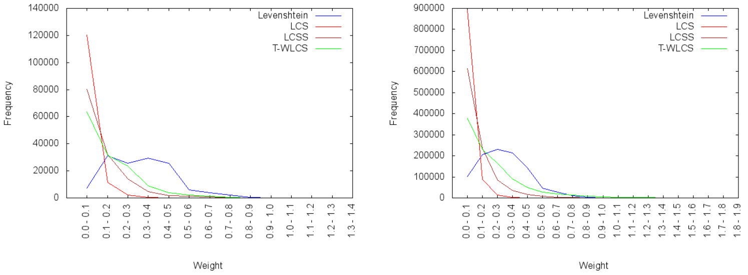

Fig. 5. Histogram - comparison of the methods - Type C(left side), Type D(right side)

4.5 Results

Left side of Fig. 6 visualizes similarity between particular sequences of case type A. Similarity was analyzed by T-WLCS method which we choose like the best representative of the tested methods according to our research that was mentioned in the section 4.4 The structure of the graph show us that where we are able to discover which sequences are more similar than others and what are the connections between them. In the future work we will analyze these parts and we will try to find why is it happening and what it means in the real business.

Right side of Fig. 6 visualizes similarity between particular sequences of case type B. That means there is involved time parameter. We can see that there are some differences according graph on the left side of the Fig. 6. In any case the result is that time parameter have some impact to the examined structure and can show us another dependencies and clusters. That is the interesting result which is in one hand essential and expectable but in the other hand bring us to another view to the examined process that we are not able to see at the beginning of analysis. There are a lot of possibilities to future work that we can do.

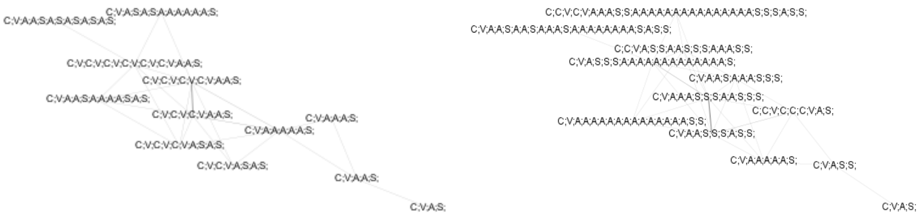
C;V;A;S;A;S;A;A;A;A;A;A;S;
C;V;A;A;S;A;S;A;S;A;S;A;S;A;S;
C;V;C;V;C;V;C;V;C;V;C;V;A;A;S;
C;V;C;V;C;V;C;V;A;A;S;
C;V;A;A;S;A;A;A;A;S;A;S;
C;V;C;V;C;V;A;A;S;
C;V;A;A;A;S;
C;V;A;A;A;A;A;S;
C;V;C;V;C;V;A;S;A;S;
C;V;C;V;A;S;A;S;
C;V;A;A;S;
C;V;A;S;
C;C;V;C;V;A;A;A;S;S;A;A;A;A;A;A;A;A;A;A;A;A;A;A;S;S;S;A;S;S;
C;V;A;A;S;A;A;S;A;A;A;S;A;A;A;A;A;A;A;A;A;S;A;S;S;
C;C;V;A;S;S;A;A;S;S;S;A;A;A;S;S;
C;V;A;S;S;S;A;A;A;A;A;A;A;A;A;A;A;A;A;S;
C;V;A;A;S;A;A;A;S;S;S;
C;V;A;A;A;S;S;S;A;A;S;S;S;
C;C;V;C;C;C;V;A;S;
C;V;A;A;A;A;A;A;A;A;A;A;A;A;A;A;S;S;
C;V;A;A;S;S;S;A;S;S;
C;V;A;A;A;A;A;S;
C;V;A;S;S;
C;V;A;S;

Fig. 6. Left figure shows graph similarity of sequences of sequence distribution Type A (T-WLCS method), Right figure shows Graph similarity of sequences of sequence distribution Type B (T-WLCS method)

5 Conclusion and Future Work

Obtained results show that the proposed approach can be successfully used for the data mining. We can relatively easily reach many types of findings that have to be analyzed then by the customer and the reasons why something is made differently or some process is made in two different ways has to be found.

Our approach allowed us to involve time and user metadata to the examination and we were able to found many interesting results. For example the concrete person behavior can be showed and analyzed what are her/his process instances and the behavior pattern. The goal of our paper was to find out whether we can use the proposed approach to this application domain and how. We have found that the application is possible after the described adjustment and brings us an opportunity to obtain interesting results from the data logs that could not have been seen before by other methods or their execution is difficult.

We would like to continue with the extension of this approach in the future. We want to find out what types of results we can obtain by the usage of different methods, examine the methods and accustom them for the usage on different real examples. Detailed interpretation of different case studies will help us then to determine which method and what views will be used then for data mining and real process examination. The idea is to have a very good control of the process by the usage of the data about already performed process instances.

Acknowledgments.

This work was supported by the project (CZ.1.05/1.1.00/02.0070), funded by the European Regional Development Fund and the national budget of the Czech Republic via the Research and Development for Innovations Operational Programme, by the project New creative teams in priorities of scientific research, reg. no. CZ.1.07/2.3.00/30.0055, supported by Operational Programme Education for Competitiveness and co-financed by the European Social Fund and the state budget of the Czech Republic and by the internal grant agency of VŠB – Technical University of Ostrava, Czech Republic, under the projects no. SP2014/157

"Knowledge modeling, simulation and design of processes" and no. SP2014/154 "Complex network analysis and prediction of network object behavior".

References

1. W.M.P. Van der Aalst, A.J.M.M. Weijters, L. Maruster: Workflow Mining: Discovering Process Models from Event Logs. *Transaction on Knowledge and Data Engineering 16(9)*, 11281142, 2004.
2. W.M.P. Van der Aalst, A.J.M.M. Weijters, Workflow Mining: Process mining: A research agenda *Computers in Industry*, 231-244, 2004.
3. Van Dongen, B.F., De Medeiros, A.K.A., Verbeek, H.M.W., Weijters, A.J.M.M., Van Der Aalst, W.M.P. 2005, "The ProM framework: A new era in process mining tool support", Lecture Notes in Computer Science, pp. 444.
4. E. Esgin, P. Karagoz. Sequence alignment adaptation for process diagnostics and delta analysis, *8th International Conference HAIS*, 191-201, 2013.
5. A. Guo and H. Siegelmann. Time-Warped Longest Common Subsequence Algorithm for Music Retrieval. *In 5th International Conference on Music Information Retrieval ISMIR*, 258261, 2004.
6. D. Gusfield, Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. *Cambridge University Press*, 2008.
7. D. S. Hirschberg. Algorithms for the Longest Common Subsequence Problem. *J., ACM*, 24:664675, October 1977.
8. Jochen De Weerdt, Annelies Schupp, An Vanderloock, Bart Baesens, Process Mining for the multi-faceted analysis of business processesA case study in a financial services organization, Computers in Industry, Volume 64, Issue 1, January 2013, Pages 57-67, 2012
9. T. Kocyan, J. Martinovič, P. Dráždilová, K. Slaninová, Searching Time Series Based On Pattern Extraction Using Dynamic Time Warping. *In Dateso, Pisek, Czech Republic*, 129-138, 2013.
10. M. Muller. Information Retrieval for Music and Motion. *Springer*, 2007.
11. X. Shi and C. C. Yang. Mining Related Queries from Search Engine Query Logs. *In 15th International Conference on World Wide Web*, 943944, NY, USA, 2006.
12. K. Slaninová, T. Kocyan, J. Martinovič, P. Dráždilová, V. Snášel. Dynamic Time Warping in Analysis of Student Behavioral Patterns. *In Dateso 2012, Zernov, Czech Republic*, Vol. 837, 49-59, 2012.
13. K. Slaninová, J. Martinovič, T. Novosad, P. Dráždilová, L. Vojáček, V. Snášel. Web Site Community Analysis Based on Suffix Tree and Clustering Algorithm. *In IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT, IEEE Computer Society*, 110-113, 2011.
14. K. Slaninová, J. Martinovič, R. Šperka, P. Dráždilová. Extraction of Agent Groups with Similar Behaviour Based on Agent Profiles. *In 12th IFIP TC8 International Conference on Computer Information Systems and Industrial Management, CISIM*, Springer-Verlag, Vol. 8104, 348-357, 2013.
15. J. Štolfa, M. Kopka, S. Štolfa, O. Koberský, V. Snášel. An Application of Process Mining to Invoice Verification Process in SAP, In 4th International Conference on Innovations in Bio-Inspired Computing and Applications, IBICA 2013, 61-74, 2014.