

Annotating sBPMN Elements with their Likelihood of Occurrence

Tobias Weller and Maria Maleshkova

Institute for Applied Informatics and Formal Description Methods, Karlsruhe
Institute of Technology (KIT),
Englerstr. 11, 76133 Karlsruhe, Germany
{tobias.weller,maria.maleshkova}@kit.edu
<http://www.aifb.kit.edu>

Abstract. Process Mining is a research discipline that aims to analyze business processes based on event logs. The event logs are among others used to create models for predicting the next activity of a given process instance. Existing models use Bayesian Networks or Markov Chains to predict the next activity in a workflow. These models require knowledge about the occurrence of activities in the business process, which is usually based on expert knowledge or based on previous workflows from event logs. Based on previous work, we will i) represent a business process in sBPMN and extend our annotation tool to ii) compute the likelihood of occurrence of activities in a business process and check for stochastic dependency in a process and iii) use the generated knowledge to annotate the business process.

Keywords: Process Models, sBPMN, Markov Chain, Annotation

1 Introduction

A very common research topic in Process Mining is the prediction of next activities in a business process. The recommendation of next activities can either be based on a schema, if a target process has been defined, on statistical methods, which exploits knowledge from past events, stored in a repository, or as a hybrid system, both in combination. The persons, involved in the process, get recommendations from a system, which next activities they should perform. Often are those recommender systems out to optimize a certain optimization function, which aims at decreasing the mortality of patients, runtime of business process or increase the satisfaction of involved persons. However, it is often very hard to predict next activities, because the underlying data is very heterogeneous and different linguistic explanations for similar activities lead to fault predictions. In addition, if no target process is defined, then the number of process variations might be high due to different opinions in the process execution by various persons. This is among others given in the medical domain. Various physicians might have different opinions and follow different best practices. This leads to a high number of process variations for similar processes. A further impact in predicting next activities is the various number of influences that comes in.

Semantics and background knowledge might improve the results of predicting next activities in a process. So far, including semantics and background knowledge, has not been considered deeply. Often, only the sequence of activities has been considered. However, there is semantics hidden in the log files of processes that can be used to improve statistical methods to predict next activities. Often occurrence of activities depend on each other.

Our previous works focused on finding correlations in meta-information of a business process. Now we are interested in finding dependencies in the workflow of a business process. Our aim is to compute the likelihood of the next activity and therefore the likelihood of a certain outcome and workflow. We will also check if events in a business process are stochastically independent. This knowledge can be used to compute the likelihood of different workflows and predicting the next activity. By annotating the business process with the likelihoods of their occurrences, a Markov chain can easily be created [6]. We will use the business process depicted in figure 1 as consistent example.

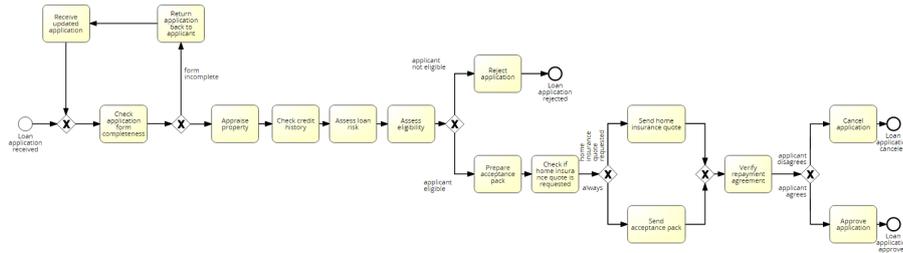


Fig. 1. Considered business process for annotating activities with their likelihood of occurrence.

Based on previous work, we will represent a business process in sBPMN and i) extend our annotation tool to ii) compute the likelihood of occurrence of activities in a business process and check for stochastic dependency in a process and iii) use the generated knowledge to annotate the business process.

In summary, this paper makes the following contributions: 1. Compute the likelihood of occurrence of activities in a business process and check for stochastically dependencies. 2. Annotate the sBPMN with the generated knowledge.

The remainder of this paper is structured as follows: the following section (Section 2) introduces the methods used to compute the likelihoods and stochastic dependencies. Section 3 demonstrates the practical applicability of our solution by realizing our approach with open-source process data and evaluating the added value to recommend next activities and computing the likelihood of certain workflows. Related work is described in section 4. We sum up our contributions and provide conclusions and future work in section 5.

2 Material and Methods

Our aim is to annotate activities of a business process with their corresponding likelihood of occurrence. This knowledge can be used to compute the likelihood of a certain outcome, workflow and use this knowledge in predicting the next activity. We start from the premise that a target process is already defined in Business Process Model and Notation (BPMN). However, because we want to exploit semantic relationship and meta-information in future, we are interested in including the semantic information into BPMN. Semantic Business Process Model and Notation (sBPMN) allows to add meaning in form of meta-information and background knowledge to each process elements. The result is a machine-readable format, which allows for reasoning on the process description. Therefore, we will use sBPMN to model the processes, because it allows to add semantics to business processes, which is used to describe the information of the business processes. Hence, a first step is to transform a given business process in BPMN into sBPMN. For this purpose, we will use previous work that automatically transforms a BPMN process in the standard format BPMN 2.0 XML by OMG [1] into sBPMN. There are already many ontologies for BPMN 2.0 available [2–5] that allows to capture the semantics in processes and include meta-information. However, not all ontologies are online available and follow the latest BPMN 2.0 version. We have to pick a suitable one. By using an ontology, we can easily add annotations to the BPMN elements.

Critical issues in a process are usually branches in which involved persons have to decide which way to pick. If no decision criteria is given, then a very basic approach is to assume a Laplace’s probability space for the next activity. This means that the likelihood of occurrence of each previous activity is the same (uniform distribution over the following activities). We exemplarily depicted in figure 2 the likelihoods of the next activity of the branches by using a Laplace’s probability space assumption. For all other previous activities, without branches, is the likelihood 1.

However, the assumption of a Laplace’s probability space is for recommending next activities in a real-world scenario not sufficient. To improve our model, we will compute the likelihood of occurrence by using existing process instances from a repository. For this purpose, we use historic data from a repository to compute the likelihood of the next activities. For sequences of activities, the next activity is obviously. Therefore, the likelihood of the next activity is 1. So crucial parts of the process are the branches for which multiple activities can follow. In our example those are the activities after *Check application form completeness*, *Assess eligibility*, *Check if home insurance quote is requested* and *Verify repayment agreement*.

By using the computed likelihoods, one can easily use them to forecast the next activity and to model a Markov chain. The assumption of Markov chains is that the next state depends only on the current state. Which was calculated above. Therefore, we can use to predict the outcome of a workflow and the likelihood of a certain workflow by using the Markov chain assumption. However, the outcome or occurrence of activities might dependent not only on the current

activities or states but on previous activities. E.g. one might think that the occurrence of returning the application back to the applicant due to an incomplete form might lead to a higher likelihood of rejecting the application. In order to check such dependencies, we will check for stochastically dependencies of activities in the process. Two events A and B are stochastically independent if the following applies:

$$P(A) \cdot P(B) = P(A \cap B)$$

3 Evaluation

We used a free available data set from BPI Challenge 2012 [7]. The log files of each process instance were created synthetically. We used the parallel target process, which contains 10,000 process instances. Meta-information, except for time-stamps, are not given. We described the target process by using Cognitive Process Designer [8, 9]. This tool is an extension to Semantic MediaWiki (SMW) [10] that allows for capturing BPMN diagrams, meta-information about the activities and describe the information semantically. Semantic MediaWiki is a powerful collaborative knowledge management system, using the MediaWiki engine and allowing for capturing information in a structured way. The captured information are stored by using RDF as standard format and can therefore be queried. By using an appropriate ontology, the target process is available in sBPMN and can easily be enriched with further information. We used an ontology published by DKM [2]. Afterwards, we uploaded the process instances into the SMW, in which the Cognitive Process Designer runs. We linked the process instances to the activities of the target process. As discussed in section 2, a Laplace's probability space might be not applicable in real-world scenarios. Therefore, we used a bash script, which counts for every occurrence of an activity the likelihood of appearance of the next activity. The fact that the business process is stored in RDF facilitated this step due to the fact that the number of activities could easily be queried. We applied this on the data set and received the results depicted in figure 2. The likelihood for sequences of activities is always one and not shown on the graphic due to reasons of clarity. Likelihoods of activities, which are not connected by a direct edge, is zero. Therefore, we only depicted the likelihoods of the next decision at decision points, because these are the crucial parts of the process. We calculated the occurrence of an activity in general as well but not depicted it on the graph.

To comprehend the application, we will show two examples in the following. In total 1,070 times were the Activity *Return application back to applicant* chosen as next activity after *Check application form completeness* and 10,000 times *Appraise property*. This leads to the following likelihoods:

$$P(X_t = \text{Appraise Property} | X_{t-1} = \text{Check application}) = \frac{1070}{11070} = 0.0967$$

$$P(X_t = \text{Return application} | X_{t-1} = \text{Check application}) = \frac{10000}{11070} = 0.9033$$

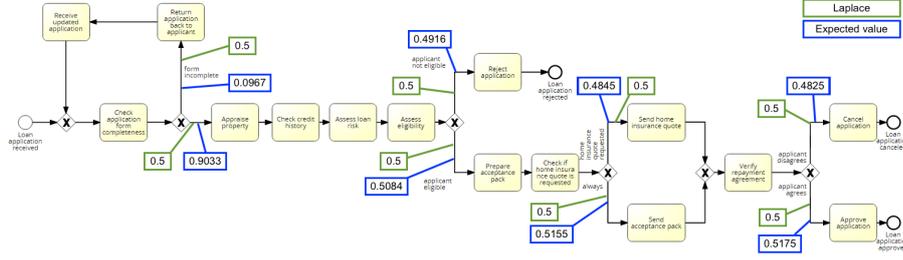


Fig. 2. Annotation of workflow by using Laplace approach and expected value for Markov Chains of next activity on decision points of the considered target process of the BPI Challenge 2012.

For the next decision point (*Asses eligibility*), we computed the following likelihoods:

$$P(X_t = \text{Reject application} | X_{t-1} = \text{Asses eligibility}) = \frac{4916}{10000} = 0.4916$$

$$P(X_t = \text{Prepare acceptance pack} | X_{t-1} = \text{Asses eligibility}) = \frac{5084}{10000} = 0.5084$$

The bash script computed stochastically dependency between activities as well. Therefore, it checks if there might be dependencies between the occurrence of certain activities exist. For instance, one might think of a stochastic dependency if an application were returned due to an incompleteness of the form and a rejection. Therefore we checked the stochastically dependency between these two events.

$$P(\text{Return application}) \cdot P(\text{Reject application}) = \frac{963}{10000} \cdot \frac{4916}{10000} = 0.0473$$

$$P(\text{Return application} \cap \text{Reject application}) = \frac{486}{10000} = 0.0486$$

Because $0.0473 \approx 0.0486$, we can assume that the events *Return application* and *Reject application* are stochastically independent. We checked other combinations of events for stochastically dependency as well. E.g. we checked if the events *Return application* and *Cancel application* are stochastically independent.

$$P(\text{Return application}) \cdot P(\text{Cancel application}) = \frac{963}{10000} \cdot \frac{2453}{10000} = 0.0236$$

$$P(\text{Return application} \cap \text{Cancel application}) = \frac{223}{10000} = 0.0223$$

Due to the fact that $0.0236 \approx 0.0223$, we assume that there is no dependency.

We performed tests by using a sampling set of 9,000 workflows for computing the likelihoods of the next activities and predicted the next activities by using the likelihoods. As assumed, because we could not find any dependency in the workflows, the likelihoods also reflect the error rate. E.g. we computed the likelihood

of the activities, occurring after *Check application for completeness*. The result was $P(\text{Appraise Property}|\text{Check application form completeness}) = \frac{9000}{9958} = 0.9038$, respectively

$P(\text{Return application back to applicant}|\text{Check application form completeness}) = 0.0962$. Therefore, as a very basic approach, we always suggested the activity with the highest likelihood for the evaluation set (1,000 workflows). In this case, the error rate was 0.1007. Which is the proportion of the likelihood of the total result $((1000 \cdot 0.1007 + 9000 \cdot 0.0962) \div 10000)$.

We used the generated knowledge about the likelihood of occurrence of each event and the likelihood of the next event in our sBPMN model by including it as meta-information. We attached these information on the activities and decision nodes. By using the likelihoods and assuming a Markov chain characteristic, we can calculate the likelihood of different workflows of the process. E.g. the likelihood that the the application is returned once to the applicant and then rejected is 0.0429. This corresponds to the likelihood of this scenario given in the data (435 times occurred this workflow¹). The respective likelihood that the application form is returned two times to the applicant and then rejected is 0.0042, which also corresponds to the occurred likelihood in the data (47 times occurred this workflow²).

4 Related Work

Our approach is addressed by roughly three kinds of work: 1) Match BPMN process to sBPMN, 2) computing likelihoods for next activities based on historic data and 3) annotating business processes with the generated knowledge.

sBPMN was developed to allow for extending BPMN elements with additional information and background knowledge to enhance analysis [11, 12]. So far, existing work already addressed the transformation of BPMN into other languages like e.g. BPEL [13, 14]. Further work developed ontologies to semantically enrich BPMN in sBPMN [2, 3]. We used an existing ontology developed by DKM [2].

Another aspect that is tackled in our approach is the computation of the occurrence of activities in a process in general, but also conditioned on previous activities, based on historic workflows. Finding predictive models to forecast activities is in fact a very prominent example in Process Mining. Forecasting tools exist that on the one hand detect data attributes that influence the choices in a process [15], as well as to detect decision points and try to minimize uncertainties in a process [16]. Existing approaches forecast the next activity and activity durations in a process by using decision trees and rule induction [17, 18], regression [19] or a classification model to support prediction of activities [20] and exceptions [21]. Other approaches addressed inferring the future actions of people from noisy visual input [22]. Approaches use Markov Decision Processes

¹ This corresponds to a likelihood of 0.0435

² This corresponds to a likelihood of 0.0047

as model and try to maximize the likelihood of the training data under the maximum entropy distribution. Surveys exist to give overviews of already addressed topics in Process Mining [23, 24].

The last addressed topic is the enrichment of the business process with the generated knowledge. Current work used semantic information to increase the precision of process models [25]. Tools exist to specify annotations for business processes [26], as well as web services [27].

5 Conclusions

In this paper we present an approach to map a BPMN process into sBPMN. We used the workflows of the process to calculate the likelihood of occurrence of each activity and the likelihood of occurrence for each next activity, depending on the current activity. This generated knowledge could in turn be used to enrich the sBPMN with further information. We used the likelihood of occurrence of the activities to check for stochastically dependencies. In addition, the likelihoods were used to compute the likelihood of various workflows. Thereby, we assumed the Markov chain characteristic of the process. Future work includes combining our previous work of detecting correlations of meta-information and the outcome of this work. We will combine the detection of correlation of meta-information and dependency of the workflow to detect crucial parts of the process and detect unknown influences in the process. We suppose to indicate the criteria in decision points.

References

1. Information technology object management group business process model and notation, URL: http://www.iso.org/iso/catalogue_detail.htm?csnumber=62652 [accessed: 2016-10-04].
2. Marco Rospocher, Chiara Ghidini, Luciano Serafini. An ontology for the Business Process Modelling Notation Formal Ontology in Information Systems - Proceedings of the Eighth International Conference, FOIS2014, September, 22-25, 2014, Rio de Janeiro, Brazil, vol. 267, pp. 133-146, IOS Press, 2014
3. Natschlger, Christine: Towards a BPMN 2.0 Ontology, Business Process Model and Notation: Third International Workshop, BPMN 2011, Lucerne, Switzerland, November 2011
4. J. vom Brocke and M. Rosemann, Eds., BPMN 2.0 for Modeling Business Processes, Handbook on Business Process Management 1: Introduction, Methods, and Information Systems. Springer, 2015, ISBN: 978-3-642-45099-0.
5. W. Yao and A. Kumar, Conflexflow: Integrating flexible clinical pathways into clinical decision support systems using context and rules, Decision Support Systems, vol. 55, no. 2, 2013, pp. 499-515, 1. Analytics and Modeling for Better HealthCare 2. Decision Making in Healthcare.
6. W. R. Gilks, S. Richardson and D. Spiegelhalter. Markov chain Monte Carlo in practice. CRC press, 1995.

7. A. Maaradji, M. Dumas, M. La Rosa and A. Ostovar, Fast and Accurate Business Process Drift Detection, QUT ePrints #83013, QUT, 2015 (<http://eprints.qut.edu.au/83013>).
8. Tobias Weller, Maria Maleshkova Cognitive Process Designer - An Open-Source Tool to Capture Processes according to the Linked Data Principles The Semantic Web: ESWC 2016 Satellite Events, Springer
9. Tobias Weller, Maria Maleshkova Capturing and Annotating Processes using a Collaborative Platform Proceedings of the 25th International Conference on World Wide Web, WWW 2016, ACM, Montral, Canada, April, 2016
10. Max Vlkol, Markus Krtzsch, Denny Vrandecic, Heiko Haller, Rudi Studer Semantic Wikipedia Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, May 23-26, 2006, ACM, Mai, 2006
11. Abramowicz, Witold, Agata Filipowska, Monika Kaczmarek and Tomasz Kaczmarek. "Semantically Enhanced Business Process Modeling Notation." *Semantic Technologies for Business and Information Systems Engineering: Concepts and Applications*. IGI Global, 2012. 259-275. Web. 3 Mar. 2017. doi:10.4018/978-1-60960-126-3.ch013
12. H. Fernndez Fernndez, Elas Palacios-Gonzlez, Vicente Garca-Daz, B. Cristina Pelayo G-Bustelo, Oscar Sanjun Martnez, Juan Manuel Cueva Lovelle, SBPMN An easier business process modeling notation for business users, *Computer Standards & Interfaces*, Volume 32, Issues 12, January 2010, Pages 18-28, ISSN 0920-5489, <http://dx.doi.org/10.1016/j.csi.2009.04.006>. (<http://www.sciencedirect.com/science/article/pii/S0920548909000300>)
13. Guillaume Doux and Frdric Jouault, Jean Bzivin. Transforming BPMN process models to BPEL process definitions with ATL, In *GraBaTs 2009 : 5th International Workshop on Graph- Based Tools*, 2009
14. Saartje Brockmans and Marc Ehrig and Agnes Koschmider and Andreas Oberweis and Rudi Studer: Semantic Alignment of Business Processes, Proceedings of the Eighth International Conference on Enterprise Information Systems (ICEIS 2006), May 2006
15. A. Rozinat, W.M.P. van der Aalst Decision mining in ProM, Proc. of 4th Intl. Conf. on Business Process Management (BPM'06) (2006), pp. 420425
16. S. Subramaniam, V. Kalogeraki, D. Gunopulos, F. Casati, M. Castellanos, U. Dayal, M. Sayal. Improving process models by discovering decision points, *Information Systems*, 32 (7) (2007), pp. 10371055
17. C. Apte, S. Weiss. Data mining with decision trees and decision rules. *Future Generation Computer Systems*, 13 (1997), pp. 197210
18. Ceci M., Lanotte P.F., Fumarola F., Cavallo D.P., Malerba D. (2014) Completion Time and Next Activity Prediction of Processes Using Sequential Pattern Mining. In: Deroski S., Panov P., Kocev D., Todorovski L. (eds) *Discovery Science. DS 2014. Lecture Notes in Computer Science*, vol 8777. Springer, Cham
19. Xiao Liu, Zhiwei Ni, Dong Yuan, Yuanchun Jiang, Zhangjun Wu, Jinjun Chen, Yun Yang, A novel statistical time-series pattern based interval forecasting strategy for activity durations in workflow systems, *Journal of Systems and Software*, Volume 84, Issue 3, March 2011, Pages 354-376, ISSN 0164-1212, <http://dx.doi.org/10.1016/j.jss.2010.11.927>.
20. Ceci, Michelangelo and Lanotte, Pasqua Fabiana and Fumarola, Fabio and Cavallo, Dario Pietro and Malerba, Donato. Completion Time and Next Activity Prediction of Processes Using Sequential Pattern Mining, *Discovery Science: 17th International Conference, DS 2014, Bled, Slovenia, October 8-10, 2014. Proceedings, 2014*, 49–61, http://dx.doi.org/10.1007/978-3-319-11812-3_5

21. Y. Hai-*tao*, D. Bin and S. Zheng-*xiao*, "Workflow Exception Forecasting Method Based on SVM Theory," 2008 International Symposium on Computational Intelligence and Design, Wuhan, 2008, pp. 81-86., <http://dx.doi.org/10.1109/ISCID.2008.66>
22. Kitani K.M., Ziebart B.D., Bagnell J.A., Hebert M. (2012) Activity Forecasting. In: Fitzgibbon A., Lazebnik S., Perona P., Sato Y., Schmid C. (eds) Computer Vision ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol 7575. Springer, Berlin, Heidelberg
23. Indranil Bose, Radha K. Mahapatra, Business data mining a machine learning perspective, Information & Management, Volume 39, Issue 3, 20 December 2001, Pages 211-225, ISSN 0378-7206, [http://dx.doi.org/10.1016/S0378-7206\(01\)00091-X](http://dx.doi.org/10.1016/S0378-7206(01)00091-X).
24. zur Mühlen, Michael and Shapiro, Robert. Business Process Analytics, Handbook on Business Process Management 2: Strategic Alignment, Governance, People and Culture, 2010, Berlin, 137–157, http://dx.doi.org/10.1007/978-3-642-01982-1_7
25. Matthias Born, Florian Drr, and Ingo Weber. 2007. User-friendly semantic annotation in business process modeling. In Proceedings of the 2007 international conference on Web information systems engineering (WISE'07), Mathias Weske, Mohand-Sad Hacid, and Claude Godart (Eds.). Springer-Verlag, Berlin, Heidelberg, 260-271.
26. K. Hinge, A. Ghose and G. Koliadis, "Process SEER: A Tool for Semantic Effect Annotation of Business Process Models," 2009 IEEE International Enterprise Distributed Object Computing Conference, Auckland, 2009, pp. 54-63. doi: 10.1109/EDOC.2009.24
27. Andreas He, Eddie Johnston, and Nicholas Kushmerick. 2004. ASSAM: a tool for semi-automatically annotating semantic web services. In Proceedings of the 3rd International Conference on Semantic Web Conference (LNCS-ISWC'04), Sheila A. McIlraith, Dimitris Plexousaki, and Frank Van Harmelen (Eds.). Springer-Verlag, Berlin, Heidelberg, 320-334. DOI: http://dx.doi.org/10.1007/978-3-540-30475-3_23