Multi-Facet BPM: Identification, Analysis and Resolution of Resource-Intensive BPM Applications

Tom Thaler, Sharam Dadashnia, Peter Fettke, Peter Loos
Institute for Information Systems (IWi) at the
German Research Center for Artificial Intelligence (DFKI) and
Saarland University
Campus D3 2, 66123 Saarbrücken
{tom.thaler|sharam.dadashnia|peter.fettke|peter.loos}@iwi.dfki.de

Abstract

Within the last years, the information systems research discipline has been faced with more and more resource intensive application scenarios and an increasing amount of data. Taking this development into account, the paper at hand exemplarily addresses three concrete calculation scenarios in order to gain insights on how to utilize a high performance IT infrastructure to solve the corresponding problem statements. Therefore, the concept of architectural prototyping is used as a research approach. This "system under development" made it possible to develop an outstanding algorithm calculating process matches and to evaluate it with the IWi process model corpus. While the work on the two other scenarios (1) state explosion in practice and (2) process mining on Big Data is still in progress, several new interesting application scenarios could be identified.

1 Introduction

The project Multi-Facet BPM aims at addressing new challenges of resource-intensive BPM application scenarios, wherefore techniques of parallel and distributed computing as well as techniques for the handling of Big Data are necessary. In order to gain insights on how to utilize a high performance IT infrastructure, the following scenarios are arranged:

1. Study the behavior of different process model similarity measures by applying them on heterogeneous data sets. Explore the existence of similarities and node correspondences in process models from different domains.
2. Study the state-explosion problem in real applications and investigate the borders of extracting all possible traces of business process models.
3. Process Mining in terms of extracting business process models from log-files on large data foundations with several millions or even billions of records.

In order to handle the mentioned scenarios, a specific research approach is applied, which is described in section 2. Section 3 provides some further information on the different scenarios and presents the accumulated results structured by established approach, research in progress and further research directions. Section 4 provides information on the developed software, while section 5 concludes the preliminary work and gives an outlook on the follow-up project.

2 Research Approach

Within the project, the concept of architectural prototyping is used as a research approach. An architectural prototype is a learning and communication vehicle used to explore and experiment with alternative architectural styles, features and patterns in order to balance different architectural qualities [1]. The main objective is to enable the calculation of the mentioned application scenarios, which is not possible with existing tools. Thus, the architectural prototype is primarily used for getting insights that may otherwise be difficult to obtain before a system is built [1].

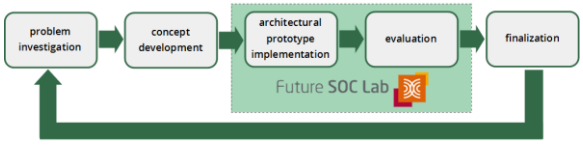

Figure 1: research approach

As shown in figure 1, the research approach is considered as a repeating cycle consisting of five phase. In the first phase, the research problem is investigated and explicated as a problem statement. Within the second phase, a concept for the solution of the problem is developed and, afterwards, implement in terms of an architectural prototype in the third phase. The implemented concept is then evaluated in phase four. At the end of an iteration, it is decided whether a further iteration is necessary or not.

The phases three and four are proceeded on the IT infrastructure provided by the HPI Future SOC Lab consisting of a dedicated blade with 24 cores, 64 GB main memory and Ubuntu as the operating system. The implementation is based on a multi-thread enabled php compilation in terms of a first architectural prototype and on java in terms of further stable implementations.

3 Calculation Scenarios and Accumulated Results

3.1 Established Approach

The first mentioned scenario aims at studying the behavior of different process model similarity measures by applying them on heterogeneous data sets. Similarity measures are necessary for the handling of large process repositories, for compliance analyses or in context of mergers and acquisitions. Calculating process similarities generally requires the availability of node matches, the assignment of node sets of one model to the corresponding node sets of another model [2]. Thereby, the investigated objects range from natural language over graphs to the execution semantics of process models. Since the generation of such matches is an optimization problem with a np-complete complexity, this can be seen as the bottleneck of the whole calculation.

However, the applied research approach made it possible to further develop a process matching algorithm, which outperforms the existing state-of-the-art algorithms in that research field and which was honored with the *Outstanding Matcher Award* at the Process Model Matching Contest 2013 [3]. Only that further development enabled the calculation of node matches between all models of the IWi process model corpus [4]. The concept of the multi-thread implementation of the algorithm (description of the algorithm in [3]) is visualized in figure 2. Two parts, namely the semantic data preparation and the binary mapping extraction could be parallelized. However, only the mapping of the SAP R/3 reference model (604 single models) on itself took about 3.5 days under maximum processor utilization of the blade and a permanent consumption of more than 32 GB main memory (see figure 3).

Within the application scenario, the matches between 2,290 single models with 63,354 nodes overall were calculated which led to more than 2 billion node comparisons and more than 2.6 million models pairs. Some interesting results are the identification of relatively high similarities between the SAP reference model and the Y-CIM reference model [5] with about 42% matched nodes and between the SAP reference model and ITIL [6] with about 36% matched nodes.

3.2 Research in Progress

3.2.1 State Explosion in Practice

The second scenario aims at investigating the theoretical state explosion problem of EPCs, which is primarily caused by the execution semantic of the OR connector. It is tried to answer the question of the relevance of that theoretical problem in real process models. Thus, also for this scenario, the IWi process model corpus is used as a data basis.

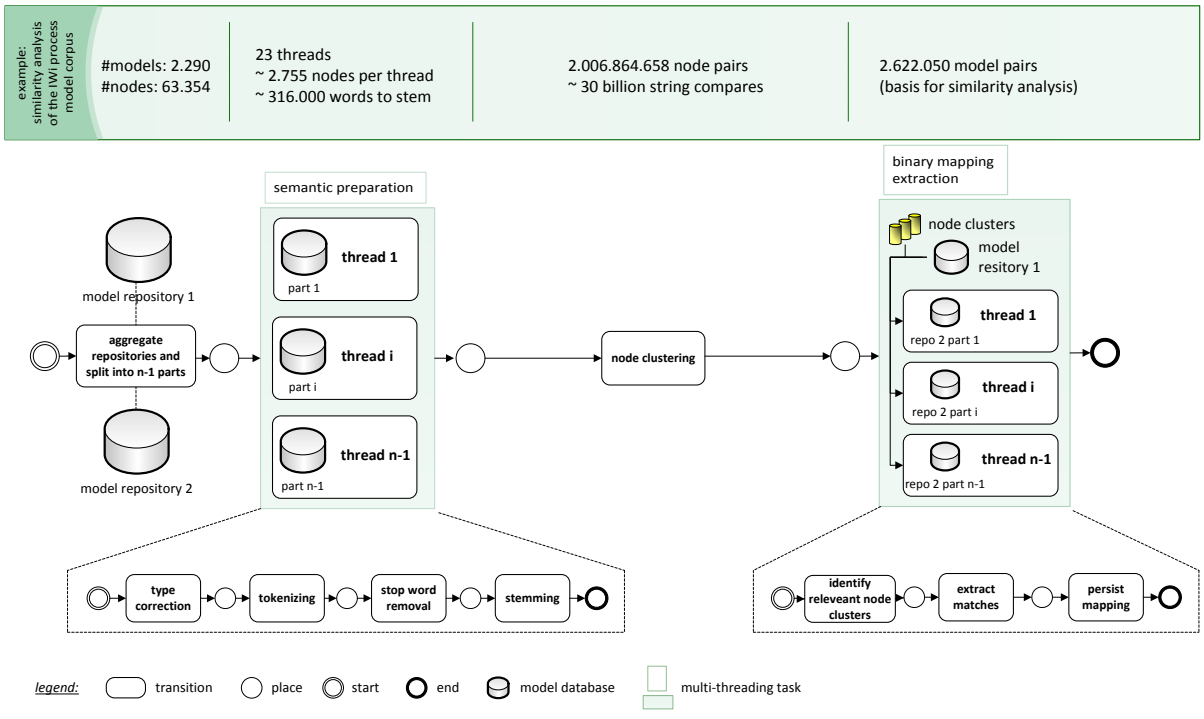

Figure 2: Multi-threaded n-ary semantic cluster matching algorithm (RefMod-Miner/NSCM)

Figure 3: Utilization while matching calculation

A concrete application scenario motivating that research is the implementation, respectively the automatic transformation, of process models to software code. The complexity of an implementation as well as the resulting software product depends (amongst others) on the number of possible execution paths.

Existing concepts addressing that topic are the graph theory in general and the token concept for EPCs and the refined process structure tree (RPST) in particular. In a first step, the token concept for EPCs [7] was implemented and extended using an architectural prototype in order to be able to handle not only models with a single entry and a single exit (SESE) but also multiple entries and multiple exits, which are very common in practice. A further algorithm calculating all possible execution paths based on reachability graphs was implemented.

Within the evaluation of the concept, 926 single models from the model corpus served as input data and it was tried to derive all possible execution paths for all models. As a result, it can be noted that for 86% of the models, all possible execution paths could be calculated within reasonable time (less than 5 minutes). The calculation aborted for 13% of the models because of time exceedance and for 1% of the models because of syntax errors within the models. It should further be noted, that the calculation nearly permanently reserved more than 10 GB of main memory, although it was processed on only one processor core.

In order to also enable the calculation of the missing 13% of the models, it is planned to (1) use the refined process structure tree for a sped up execution path derivation and to (2) identify tasks within the algorithm which can be parallelized, to develop a corresponding concept and to implement that concept. Thus, it is necessary to carry out further iterations of the mentioned lifecycle in section 2.

3.2.2 Process Mining on Big Data

The third scenario aims at developing new process mining algorithms, which are able to handle large amounts of instance data. The main objects of interest are instance logs (how to reduce the mass of data), instance cluster techniques (how to cluster the instance data in a meaningful manner – e.g. in order to generate manageable process models) and process mining techniques (how to design process mining algorithms being able to handle large log files). In that context, process mining algorithms are not only used for process discovery but also for checking the conformance of process executions to the planned processes and the enhancement of existing models with data from real execution.

This scenario is currently in the concept development phase of the presented lifecycle and focuses on the discovery of process models from large log files. It is investigated which possibilities of parallelizing the mining process do exist and which software infrastructure might be meaningful for that task. In a first step, some simple algorithms like the alpha algorithm will be implemented on the Hadoop Map-Reduce Framework to replicate recent research findings [8]. Furthermore, an expansion in terms of an analogue implementation of state-of-the-art algorithms is planned.

The evaluation scenario is covered by the log data of an android app on mobile devices. Thereby, more than 6,000 users generate more than 81 million records with more than 850,000 different tasks every month. Based on that data, it should be analyzed which usage scenarios do exist on mobile devices and whether it is possible to identify different user groups.

3.2.3 Further Research Directions

During the investigation of the mentioned scenarios, some further research directions were identified, which should briefly be introduced in the following:

- Process clustering: In context of mergers and acquisitions, it might be meaningful to cluster similar processes, e.g. in order to compare them.
- Process integration: As a follow-up step of process clustering, e.g. in order to standardize business processes, one possibility is the integration of process models which leads to a new process model aggregating all commonalities and differences of all input models.
- Inductive reference model development: Next to the traditional way of developing reference models in a deductive manner, another way is the inductive development of reference models based on existing models. The idea is to extract the best known practice and use that information to construct a new model.

- Model corpora and catalogs: As described in the previous sections, model collections like specific process model corpora or catalogues serve as adequate input data for several evaluation scenarios. Against that background, the development, analysis and the usage of such corpora are named as a further research direction as only they enable the replicability of research findings.
- Natural language processing: Natural language, e.g. in the form of node labels, process descriptions or meta-data is a very important artifact in business process management. Thus, NLP techniques are e.g. used for the identification of correspondences or for text-to-model / model-to-text transformations.

4 RefMod-Miner

As mentioned in section 2, there are two implementation stages. The first stage (php - solely command line) is primarily used for first drafts and is characterized by a trial and error implementation. This is based on that fact that, in php, types can be neglected in most cases, which leads to first results in very short time. The source code of the existing implementation is publicly available and can be downloaded at https://refmodmine.googlecode.com/svn.

The second stage is developed in Java and covers the more stable research prototype which is called RefMod-Miner. Generally approved approaches are implemented and implementations in an early state are explicitly marked as such. The RefMod-Miner as well as the corresponding documentation and exemplarily use cases are available at http://refmodminer.dfki.de.

5 Conclusion and Outlook

The project Multi-Facet BPM made a first step towards addressing new requirement regarding the need for high performance computing in the field of business process management. A concept of architectural prototyping was used as a research approach and delivered insights in context of the focused application scenarios, which may otherwise be much more difficult to obtain before a system is built.

The adaption of an already professionally approved technique for the identification of correspondences between nodes of process models led to two important results. First, only that enabled the application of the technique on a large amount of data. Without that further development, it would not be possible to analyze a large model corpus with regard to the contained similarities. Second, it allowed the collection of experiences on how to design an adequate software in order to ideally utilize a high performance IT infrastructure.

However, the investigation of the two other scenarios is still in progress. Indeed, there exists a first results in the area of exploring the state explosion problem of EPCs in practice. Nevertheless, further iterations of the architectural prototyping lifecycle are necessary for a concluding statement.

The field of process mining on Big Data is currently in the phase of concept development and will be further arranged in a follow-up project.

Another result of the project is the identification of five additional scenarios, which will be focused in a later period of HPI Future SOC Lab.

Acknowledgement

The provided high performance IT infrastructure from the HPI allowed the investigation of concrete problem fields in information systems research. The authors thank the HPI Future SOC Lab for the chance of using these resources and appreciate a continuation of the project.

The basic concepts were developed in context of the project “Konzeptionelle, methodische und technische Grundlagen zur induktiven Erstellung von Referenzmodellen (Reference Model Mining)”, which is funded by the Deutsche Forschungsgemeinschaft DFG (GZ LO 752/5-1).

References

[1] Bardram, J. E., Christensen, H. B., Hansen, K. M.: Architectural prototyping: an approach for grounding architectural design and learning, In: Proceedings of the Fourth Working IEEE/IFIP Conference on Software Architecture, IEEE, pp. 15-24, 2004.

[2] Thaler, T., Hake, P., Fettke, P., Loos, P.: Evaluating the Evaluation of Process Matching Technique, In: Leena Suhl; Dennis Kundisch (ed.). Tagungsband der Multikonferenz Wirtschaftsinformatik (MKWI2014), February 26-28, Paderborn, Germany, Universität Paderborn, pp. 1600-1612, 2014.

[3] Cayoglu, U., Dijkman, R., Dumas, M., Fettke, P., Garcia-Banuelos, L., Hake, P., Klinkmüller, C., Leopold, H., Ludwig, A., Loos, P., Mendling, J., Oberweis, A., Schoknecht, A., Sheetritt, E., Thaler, T., Ullrich, M., Weber, I., Weidlich, M.: The Process Model Matching Contest 2013, In: Business Process Management Workshops – BPM 2013 International Workshops, Beijing, China, Springer International, pp. 442-463, 2013.

[4] Thaler, T., Walter, J., Ardalani, P., Fettke, P., Loos, P.: Development and Usage of A Process Model Corpus, In: Proceedings of the 24th International Conference on Information Modelling and Knowledge Bases EJC 2014. June 3-6, Kiel, Germany, 2014.

[5] Scheer, A.-W.: Business Process Engineering - Reference Models for Industrial Enterprises. 2. ed., Berlin, Springer, 1994.

[6] Office of Government Commerce, ITIL - Service Strategy, Service Design, Service Operation, Service

Transition, Continual Service Improvement, Norwich TSO Information & Publishing Solutions, 2010

[7] Mendling, J.: Detection and Prediction of Errors in EPC Business Process Models. Doctoral Thesis, Vienna University of Economics and Business Administration. Vienna, Austria, May 2007.

[8] Evermann, J., Assadipour, G.: Big Data meets Process Mining: Implementing the Alpha Algorithm with Map-Reduce. ACM Symposium on Applied Computing, Gyeongju, Korea, 2014.