m^3 – A Behavioral Similarity Metric for Business Processes

Matthias Kunze, Matthias Weidlich, and Mathias Weske
{matthias.kunze, matthias.weidlich, mathias.weske}@hpi.uni-potsdam.de

Hasso Plattner Institute at the University of Potsdam
Prof.-Dr.-Helmert-Strasse 2-3, 14482 Potsdam

Abstract. With the increasing uptake of business process management, companies maintain large scale process repositories consisting of hundreds or thousands of process models. So far, discovery within these repositories is limited to free text search or folder navigation. In a separate stream of research, similarity measures were introduced to get a better understanding of the relationships between process models. Unfortunately, calculating such similarity is complex, so that these techniques cannot be used in large process model repositories, where they would be most valuable.
To overcome this issue, we introduce the m^3-metric, which is based on behavioral profiles that provide an abstraction on the detailed behavior of processes. This metric can be computed efficiently and enables tree based similarity search within large process model repositories.

1 Introduction

In recent years we saw large business process model collections grow in many organizations, whereas the effective management of such repositories requires efficient capabilities to find process models among hundreds or thousands of candidate models. The question of similarity between process models has been thoroughly studied. Still, existing approaches do not scale well in computation complexity, nor do they address transitivity, which is essential for efficient similarity search.

Similarity metrics provide such a property and significantly increase search performance, as we showed for process model structures, i.e., the graph edit distance, in [7]. In this paper we address behavioral aspects of processes and present the m^3-metric: A metric based on behavioral profiles that provides a similarity ranking of process models relative to a given query model and can be employed in metric similarity search methods, cf. [14]. Behavioral profiles focus on ordering relations between pairs of activities in a process model. While this notion abstracts from the actual behavior of a process, it is computed efficiently [11]. Approaches that take the complete state space of a process into account, in turn, suffer from exponential complexity due to the state space explosion problem.

The remainder of this work is structured as follows: In Section 2 we present previous work related to the topic of process model similarity and searching, while

Section 3 introduces formal concepts for the m^3-metric. In Section 4 we show how the aforementioned metric is constructed from behavioral profile relations and present its rationale by means of an illustrative example, before we conclude this work and give an outlook on future studies in Section 5.

2 Related Work

Similarity of process models has been addressed from various angles. An overview of linguistic, structural, and behavioral measures used for similarity search of process models can be found in [4]. Measures for structural similarity, e.g., the one based on the graph edit distance [3], neglect common behavior expressed in a different syntax when comparing process models. Modeling a loop with a loop activity in BPMN or with a control flow cycle would, therefore, impact on structural similarity of process models in a negative manner. Measures for behavioral similarity are insensitive to such syntactical differences. They may be based directly on the sets of possible traces of process models, e.g., by computing the intersection of traces of two models. In order to get a more fine-granular measure, an n-gram representation of the sets of traces may be used to judge on similarity [12]. Other approaches advocate the application of causal footprints to approximate the behavior and to measure similarity of process models [10]. Still, these approaches are computationally hard, so that recent techniques aim at a multi-step approach that narrows the search space in a step-wise manner [13]. We avoid such problems as behavioral profiles are computed efficiently for a broad class of process models. A behavioral abstraction close to the behavioral profile has been applied for matching BPEL process definitions [5]. However, the approach is restricted to BPEL processes and transitivity aspects of the proposed measures are not discussed.

In traditional databases, data is generally made up of simple structures and attribute data—indexing techniques have been very successfully elaborated on and implemented. However, for complex data, such as process models, these techniques are not applicable, because no intrinsic ordering exists among data objects and mapping them to simple values, i.e., hashing, is not meaningful. Similarity search addresses this field where nothing but pairwise distances between data objects can be measured [14]. This concept requires the distance—or dissimilarity—of two objects to be a proper metric, and thus to provide transitivity. By that, it becomes possible to predict or at least constrain the distance of a pair of data objects, if one knows the respective pairwise distances of these data objects to a third one. Several indexing techniques have been developed [2,6]. However, the above process model similarity measures have not been shown to provide proper metrics.

3 Background

This section introduces the background of our work in terms of the characteristics of a distance metric, a formal model, and the concept of a behavioral profile.

3.1 Distance Metric

To efficiently[1] search within a space of given objects, it is necessary to partition that space and exclude some of the partitions from exhaustive search. Partitioning is relatively easy for objects whose features can be mapped to vectors, i.e., in coordinate spaces. However, such a representation cannot be generally assumed, in particular for process behavior or graph structures, cf. [7]. However, in metric spaces—a generalization of coordinate spaces—nothing but a distance with certain properties is required to partition the space, the notion of such a distance is a metric [14].

Definition 1 (Distance Metric). *A metric space is a pair* $\mathcal{S} = (\mathcal{D}, d)$ *where* $\mathcal{D}$ *is the domain of objects and* $d : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ *is a* metric*, i.e., a distance function with the following properties:*

- *symmetry:* $\forall o_i, o_j \in \mathcal{D} : d(o_i, o_j) = d(o_j, o_i)$
- *nonnegativity:* $\forall o_i, o_j \in \mathcal{D}, o_i \neq o_j : d(o_i, o_j) > 0 \bigwedge \forall o_i \in \mathcal{D} : d(o_i, o_i) = 0$
- *triangle inequality:* $\forall o_i, o_j, o_k \in \mathcal{D} : d(o_i, o_k) \leq d(o_i, o_j) + d(o_j, o_k)$

Particularly, the triangle inequality states that every pair of distances between three objects is larger than the remaining. This allows deriving minimum and maximum bounds for the distance of two points, if their respective distances to another point are given, and thus partitioning the search space.

3.2 Process Models

A process model is based on a graph containing activity nodes and control nodes. It captures the commonalities of most process description languages.

Definition 2 (Process Model). *A* process model *is a tuple* $P = (A, s, C, N, F, T)$ *where:*

- A *is a finite non-empty set of activity nodes,*
- C *is a finite set of control nodes,*
- $N = A \cup C$ *is a finite set of nodes with* $A \cap C = \emptyset$*,*
- $F \subseteq N \times N$ *is the flow relation, such that* (N, F) *is a connected graph,*
- $\bullet n = \{n' \in N | (n', n) \in F\}$ *and* $n\bullet = \{n' \in N | (n, n') \in F\}$ *denote direct predecessors and successors, we require* $\forall\ a \in A : |\bullet a| \leq 1 \wedge |a \bullet| \leq 1$*,*
- $s \in A$ *is the only start node, such that* $\bullet s = \emptyset$ *and* $\forall\ n \in N : s\ F^*\ a$ *with* F^* *as the reflexive transitive closure of* F*,*
- $T : C \to \{and, xor\}$ *associates each control node with a type.*

We assume trace semantics for process models. The behavior of a process model $P = (A, s, C, N, F, T)$ is a set of *traces* $\mathcal{T}_P$. It comprises a set of lists of the form $\sigma = \langle s, a_1, \ldots, a_n \rangle$ with $n > 0$, $n \in \mathbb{N}$, $a_i \in A$ for all $0 < i \leq n$, which represent the execution order of activities. These traces follow on common Petri net-based formalizations [9].

[1] An efficient algorithm is one that avoids examining every point in the set.

3.3 Behavioral Profiles

A behavioral profile captures behavioral characteristics of a process model by three relations between pairs of activity nodes. These relations are based on the notion of *weak order*. Two activities of a process model are in weak order, if there exists a trace in which one activity occurs after the other.

Definition 3 (Weak Order Relation). *Let $P = (A, s, C, N, F, T)$ be a process model and $\mathcal{T}_P$ its set of traces. The* weak order relation $\succ_P \subseteq A \times A$ *contains all pairs (x, y), such that there is a trace $\sigma = n_1, \ldots, n_m$ in $\mathcal{T}_P$ with $j \in \{1, \ldots, m-1\}$ and $j < k \leq m$ for which holds $n_j = x$ and $n_k = y$.*

Based on the weak order relation, the behavioral profile is defined as follows.

Definition 4 (Behavioral Profile). *Let $P = (A, s, C, N, F, T)$ be a process model. A pair $(x, y) \in A \times A$ is in one of the following relations:*
- *The* strict order relation $\rightsquigarrow_P$*, if $x \succ_P y$ and $y \not\succ_P x$.*
- *The* exclusiveness relation $+_P$*, if $x \not\succ_P y$ and $y \not\succ_P x$.*
- *The* interleaving order relation $||_P$*, if $x \succ_P y$ and $y \succ_P x$.*

The set $\mathcal{B}_P = \{\rightsquigarrow_P, +_P, ||_P\}$ of all three relations is the behavioral profile *of P.*

We illustrate the relations of the behavioral profile for the BPMN model in Fig. 1. It holds $A \rightsquigarrow D$ as both activities are ordered if they occur together in a trace and $B||C$ due to the concurrent execution of both activities. An activity is either exclusive to itself (e.g., $A + A$ in Fig. 1) or in interleaving order to itself. Further details on behavioral profiles can be found in [11], which also shows how a behavioral profile of a process model is computed in polynomial time to the size of the model under the assumption of soundness. Soundness is a correctness criterion that guarantees the absence of behavioral anomalies [1].

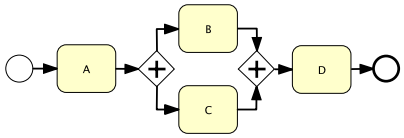

Fig. 1. Example BPMN model a

4 Construction of the m³-Metric

We assume two process models P and Q to be similar, if they expose a common share of behavior, i.e., they have a common set of activities that yield equal behavioral profiles: $(\rightsquigarrow_P \cap \rightsquigarrow_Q) \cup (+_P \cap +_Q) \cup (||_P \cap ||_Q) \neq \emptyset$. The larger this overlap of behavioral profiles is, the more similar two process models are. We quantify this overlap by means of the established Jaccard similarity coefficient for the similarity of two sets: $sim(A, B) = \frac{|A \cap B|}{|A \cup B|}$. If two sets of behavioral profile relations consist of the same pairs, they are equal, i.e., their similarity is 1. If two behavioral profile relations have no common pairs, their similarity coefficient is 0. From the relations of the behavioral profile we propose three individual similarity coefficients:

Exclusiveness Similarity captures the amount of exclusiveness, i.e., pairs of activities that must not occur together, shared by the two models, $s_{+}(P,Q) = \frac{|+_P \cap +_Q|}{|+_P \cup +_Q|}$.

Strict Order Similarity quantifies to which degree two processes expose an overlap in their order dependencies for pairs of activities, $s_{\rightsquigarrow}(P,Q) = \frac{|\rightsquigarrow_P \cap \rightsquigarrow_Q|}{|\rightsquigarrow_P \cup \rightsquigarrow_Q|}$.

Interleaving Order Similarity accounts for the observation that parallel execution of activities covers also sequential execution of the same activities in any order, i.e., activities that are executed in parallel can also be executed in a certain sequence and the according traces are therefore considered similar. $s_{||}(P,Q) = \frac{1}{2} \cdot \left(\frac{|(\rightsquigarrow_P \cup ||_P) \cap \rightsquigarrow_Q|}{|\rightsquigarrow_P \cup ||_P \cup \rightsquigarrow_Q|} + \frac{|\rightsquigarrow_P \cap (\rightsquigarrow_Q \cup ||_Q)|}{|\rightsquigarrow_P \cup \rightsquigarrow_Q \cup ||_Q|} \right)$.

A distance metric expresses a dissimilarity of two objects. Analogously, there exists a set distance that is constructed from the Jaccard similarity coefficient which has been proven to be a metric [8]: $d(A,B) = 1 - sim(A,B)$. Thus, each of the aforementioned similarity measures translates into a single distance metric. Through weighted summation of these three single metrics, we can compose them into one (thus the name m^3-metric). This composition preserves the properties of a metric.

Definition 5 (m^3-metric). *Let P and Q be two process models and $s_{+}(P,Q)$, $s_{\rightsquigarrow}(P,Q)$, and $s_{||}(P,Q)$ the similarity metrics based on behavioral profiles. Then, the m^3-metric is defined as*

$$m^3(P,Q) = 1 - \sum_i w_i \cdot s_i(P,Q)$$

with $i \in \{+, \rightsquigarrow, ||\}$ and weighting factors $w_i \in (0,1)$ such that $\sum_i w_i = 1$.

To illustrate this metric consider the sample processes, Fig. 1-3. The relations of the behavioral profile for these models are summarized in Table 1. We chose the following weights to demonstrate the m^3-metric: $w_{+} = 0.5, w_{\rightsquigarrow} = 0.3, w_{||} = 0.2$. Here, we understand exclusiveness as the strictest criterion and thus give it the highest weight to penalize violations thereof. Interleaving order offers the greatest flexibility and thus is considered the weakest criterion, which is why it receives the smallest weight.

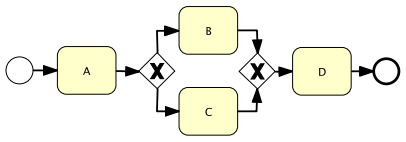

Fig. 2. Example BPMN model b

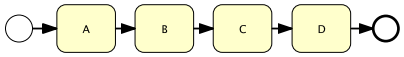

Fig. 3. Example BPMN model c

Table 1. Relations of the behavioral profile for the example process models

	Model a, Fig. 1	Model b, Fig. 2	Model c, Fig. 3
$+$	{(A,A), (B,B), (C,C), (D,D)}	{(A,A), (B,B), (B,C), (C,C), (D,D)}	{(A,A), (B,B), (C,C), (D,D)}
$\rightsquigarrow$	{(A,B), (A,C), (A,D), (B,D), (C,D)}	{(A,B), (A,C), (A,D), (B,D), (C,D)}	{(A,B), (A,C), (A,D), (B,C), (B,D), (C,D)}
$\|\|$	{(B,C)}	$\emptyset$	$\emptyset$

Building the metric space of behavioral profiles $\mathcal{M} = (\mathcal{B}, m^3)$ for the three example process models and computing the m^3-distances, we get $m^3(a, b) = 0.117$ and $m^3(b, c) = 0.183$. According to our metric, the behavioral distance between models a and b is smaller than the one between models b and c.

Since m^3 is a metric, it features the triangle inequality, which allows us to bound the distance of models a and c without actually computing it. Based on Def. 1, $|m^3(a, b) - m^3(b, c)| \leq m^3(a, c)$ (lower boundary) and $|m^3(a, b) + m^3(b, c)| \geq m^3(a, c)$ (upper boundary), i.e., $0.066 \leq m^3(a, c) \leq 0.3$. This approximation is confirmed by the actual computed value, which is $m^3(a, c) = 0.067$. The computed distances also comply with our perception of the behavioral similarity of the sample process models. Since possible traces of a cover the traces of c, due to the parallel branch, these two models are more similar (less distant) to each other than a and b, and b and c respectively.

5 Conclusion

Efficient similarity search requires a distance notion that obeys to certain properties: It must be a proper metric. We proposed a metric that allows comparing and searching process models with behavioral aspects in mind, based on the concept of behavioral profiles. These profiles are computed efficiently for a broad class of process models [11]. We explained that metric with a simple example.

The presented metric is our first attempt to investigate similarity of process models in terms of behavioral profiles. In future work, we shall address the metric, identify and rank further similarity coefficients, and construct a more sophisticated metric that is substantiated through exhaustive experiments, e.g., a regression analysis. The expressiveness of such a metric shall be compared to a reference model collection that has been evaluated by business process experts. Further, we will address the suitability of this improved metric in similarity search, as it is vital for a metric to be well discriminating in order to enable efficient searching with confident results.

References

1. W.M.P. van der Aalst. Workflow verification: Finding control-flow errors using petri-net-based techniques. In *BPM*, volume 1806 of *LNCS*, pages 161–183, 2000.
2. Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. Searching in Metric Spaces. *ACM Comput. Surv.*, 33(3):273–321, 2001.
3. Remco M. Dijkman, Marlon Dumas, and Luciano García-Bañuelos. Graph matching algorithms for business process model similarity search. In Umeshwar Dayal, Johann Eder, Jana Koehler, and Hajo A. Reijers, editors, *BPM*, volume 5701 of *Lecture Notes in Computer Science*, pages 48–63. Springer, 2009.
4. Marlon Dumas, Luciano García-Bañuelos, and Remco M. Dijkman. Similarity search of business process models. *IEEE Data Eng. Bull.*, 32(3):23–28, 2009.
5. Rik Eshuis and Paul W. P. J. Grefen. Structural matching of bpel processes. In *ECOWS*, pages 171–180. IEEE Computer Society, 2007.
6. Gisli R. Hjaltason and Hanan Samet. Index-driven similarity search in metric spaces (survey article). *ACM Trans. Database Syst.*, 28(4):517–580, 2003.
7. Matthias Kunze and Mathias Weske. Metric Trees for Efficient Similarity Search in Process Model Repositories. In *Proceedings of the 1st International Workshop on Process in the Large (IW-PL '10)*, Hoboken, NJ, September 2010.
8. Alan Lipkus. A Proof of the Triangle Inequality for the Tanimoto Distance. *Journal of Mathematical Chemistry*, 26:263–265, 1999. 10.1023/A:1019154432472.
9. Niels Lohmann, Eric Verbeek, and Remco M. Dijkman. Petri net transformations for business processes - a survey. *T. Petri Nets and Other Models of Concurrency*, 2:46–63, 2009.
10. Boudewijn van Dongen, Remco Dijkman, and Jan Mendling. Measuring Similarity between Business Process Models. In *Advanced Information Systems Engineering*, volume 5074 of *Lecture Notes in Computer Science*, pages 450–464. Springer Berlin / Heidelberg, 2008.
11. Matthias Weidlich, Jan Mendling, and Mathias Weske. Efficient consistency measurement based on behavioural profiles of process models. *IEEE Transactions on Software Engineering*, 2010. To appear.
12. Andreas Wombacher. Evaluation of technical measures for workflow similarity based on a pilot study. In Robert Meersman and Zahir Tari, editors, *OTM Conferences (1)*, volume 4275 of *Lecture Notes in Computer Science*, pages 255–272. Springer, 2006.
13. Zhiqiang Yan, Remco M. Dijkman, and Paul Grefen. Fast business process similarity search with feature-based similarity estimation. In Robert Meersman, Tharam S. Dillon, and Pilar Herrero, editors, *OTM Conferences (1)*, volume 6426 of *Lecture Notes in Computer Science*, pages 60–77. Springer, 2010.
14. Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. *Similarity Search: The Metric Space Approach.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.