Behavioral Similarity
A Proper Metric
Abstract

A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.
Abstract

A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.
A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.
Abstract

A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.

What is a proper metric?

What can it be used for?

How does it work?
Abstract

A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.

What is a proper metric?

What can it be used for?

How does it work?

How well does it work?
Abstract

What is a proper metric?

A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.

How does it work?

What can it be used for?

How well does it work?
Search in Process Repositories

• **large process model repositories** require meaningful means for search
  • find similar models to given model
  • identify duplicates

• Similarity Search: Find all process models within a *given proximity* to a given *query model*. 
Efficient Similarity Search

- $t_{seach} = t_{comparison} \cdot |comparisons|$

- data preprocessing and reduction reduce $t_{comparison}$

- index data structures and algorithms reduce $|comparisons|$

- traditional indexing: requires transitive relation, e.g., ordering or coordinates
“Proper” Metric

• A metric $d$ is distance function, that is symmetric, non-negative and obeys the triangle inequality.

• No coordinates or ordering required!
“Proper” Metric

• A metric $d$ is distance function, that is symmetric, non-negative and obeys the triangle inequality.

• No coordinates or ordering required!
“Proper” Metric

• A metric $d$ is distance function, that is symmetric, non-negative and obeys the triangle inequality.

• No coordinates or ordering required!
“Proper” Metric

• A metric \( d \) is distance function, that is symmetric, non-negative and obeys the triangle inequality.

• No coordinates or ordering required!

\[
\forall o_i, o_j, o_k \in \mathcal{D} : d(o_i, o_k) \leq d(o_i, o_j) + d(o_j, o_k)
\]

cannot be larger than \( a + b \)
Similarity Search with a Metric

• Given a partition that includes a set of models with \( p \) at its center and a search query \( q \).
• Given a partition that includes a set of models with \( p \) at its center and a search query \( q \).
Exclusion

• No model in the partition of \( p \) can match \( q \).

\[
r(q) + r(p) < d(p, q)
\]

\( r(p) \) and \( r(q) \) are radii of the spheres around \( p \) and \( q \) respectively. The distance \( d(p, q) \) is the distance between \( p \) and \( q \).
• No model in the partition of $p$ can match $q$.

\[ r(q) + r(p) < d(p, q) \]
Intersection

• Partition of \( p \) cannot be excluded and must be searched exhaustively.

\[
 r(q) + r(p) \geq d(p, q)
\]
Intersection

- Partition of $p$ cannot be excluded and must be searched exhaustively.

$$r(q) + r(p) \geq d(p, q)$$
Partition of $p$ cannot be excluded and must be searched exhaustively.

\[ r(q) + r(p) \geq d(p, q) \]
Efficient Search

• If we **partition the whole repository**, we can **safely exclude partitions from search** and become efficient.
Efficient Search

• If we **partition the whole repository**, we can safely exclude partitions from search and become efficient.
Efficient Search

- If we **partition the whole repository**, we can **safely exclude partitions from search** and become efficient.
Abstract

A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.

How does it work?

How well does it work?
Behavioral Profile

- describes the **behavioral relation** between any **pair of activities** in a process model
- is an **abstraction of behavior**, based on trace semantics
Strict Order

- $A \prec B$

- in any trace of the process, in which $A$ and $B$ occur, $A$ happens before $B$
• A \leadsto B

• in any trace of the process, in which A and B occur, A happens before B
Exclusiveness Relation

- \( B + C \)

- two activities B and C never occur together in a trace of the process
Interleaving Order

- B || C
- \textbf{else}, i.e., B may happen before C and vice versa
Interleaving Order

• B || C

• else, i.e., B may happen before C and vice versa

information loss: CAUSALITY, CARDINALITY
Behavioral Profile

• describes relation between any pair of activities in a model

• abstraction – makes comparison faster

• can be (pre-) computed efficiently for sound, free-choice process models
Design of a Metric

- Jaccard Distance for sets
  \[ d = 1 - \frac{|A \cap B|}{|A \cup B|} \]
  A and B measures their “overlap”

- compare particular relations of the behavioral profiles of two process models, P and Q
  - elementary similarities (\(\to\), +, | |)
  - extended similarities
Design of a Metric

proven to be a metric by Lipkus [34]

• Jaccard Distance for sets
  \[ d = 1 - \frac{|A \cap B|}{|A \cup B|} \]
  A and B measures their „overlap“

• compare particular relations of the behavioral profiles of two process models, P and Q
  • elementary similarities (\(\rightarrow\), +, \(\mid\mid\))
  • extended similarities

[34]
Strict Order Similarity

- reward similar ordering of common activities in process models $P$ and $Q$

\[
sim_{\sim}(\mathcal{B}_P, \mathcal{B}_Q) = \frac{|\sim_P \cap \sim_Q|}{|\sim_P \cup \sim_Q|}
\]

\[\begin{align*}
\text{Diagram 1} & : A \rightarrow B \rightarrow C \\
\text{Diagram 2} & : A \rightarrow B \rightarrow C
\end{align*}\]
• reward **similar ordering** of common activities in process models $P$ and $Q$

\[
sim_{\sim}(\mathcal{B}_P, \mathcal{B}_Q) = \frac{|\sim P \cap \sim Q|}{|\sim P \cup \sim Q|}
\]

\[
sim_{\sim}(\mathcal{B}_P, \mathcal{B}_Q) = \frac{|\{(A,B),(A,C)\} \cap \{(A,B),(A,C)\}|}{|\{(A,B),(A,C)\} \cup \{(A,B),(A,C)\}|} = 1
\]
Elementary Similarities

**Exclusiveness Similarity**

- penalizes violations of exclusiveness constraints

\[ \text{sim_+}(\mathcal{B}_P, \mathcal{B}_Q) = \frac{|P \cap Q|}{|P \cup Q|} \]

**Interleaving Order Similarity**

- rewards high degree of flexibility
- e.g. loops and concurrency

\[ \text{sim_\parallel}(\mathcal{B}_P, \mathcal{B}_Q) = \frac{||P \cap Q|}{|P \cup Q|} \]
Extended Similarities

Extended Strict Order Similarity

• rewards strict order, even if in opposite directions

\[
sim_{\sim'}(\mathcal{B}_P, \mathcal{B}_Q) = \frac{|(\sim_P \cup \sim_P^{-1}) \cap (\sim_Q \cup \sim_Q^{-1})|}{|(\sim_P \cup \sim_P^{-1}) \cup (\sim_Q \cup \sim_Q^{-1})|}
\]

Extended Interleaving Order Similarity

• rewards any order unless exclusiveness is violated

\[
sim_{||'}(\mathcal{B}_P, \mathcal{B}_Q) = \frac{|(\sim_P \cup \sim_P^{-1} \downarrow_P) \cap (\sim_Q \cup \sim_Q^{-1} \downarrow_Q)|}{|(\sim_P \cup \sim_P^{-1} \downarrow_P) \cup (\sim_Q \cup \sim_Q^{-1} \downarrow_Q)|}
\]
• **Aggregate** the similarities by **weighted sum** and transform it into a metric.

• weights determined through experiments

\[ d_B(B_P, B_Q) = 1 - \sum_h w_h \cdot \text{sim}_h(B_P, B_Q) \]

\[ h \in \{+, \sim, \|, \sim', \|'\} \]
Behavioral Profile Metric

- **Aggregate** the similarities by **weighted sum** and transform it into a metric.
- Weights determined through experiments

\[
\begin{align*}
d_B(\mathcal{B}_P, \mathcal{B}_Q) &= 1 - \sum_h w_h \cdot \text{sim}_h(\mathcal{B}_P, \mathcal{B}_Q) \\
&\quad h \in \{+, \sim, ||, \sim', ||'\}
\end{align*}
\]

Proof that this is a metric can be found in the paper.
Abstract

A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.

A suitable distance function

Overlap in behavioral profile relations

Reduce number of comparisons

How well does it work?
Evaluation

• SAP Reference Model (604 EPC)

• **human assessment of similarity** among a subset of pairs provided by Dijkman et al. [2]

• identified matching nodes by string edit distance of their labels

• measured **precision** of metrics for different **recall** levels
• **single metrics**

• **aggregated metric with certain weights**
Search Performance

- built metric tree index and searched for 10 most similar process models
Search Performance

• built metric tree index and searched for 10 most similar process models

Beyond the paper
Search Performance

• built metric tree index and searched for 10 most similar process models
• saved up to $90\%$ of $|\text{comparisons}|$ compared to linear search

Beyond the paper
Limitations

- Efficiency is dependent on the distance distribution among model collection.
- Limited support for behavioral excerpts/subgraphs.
Conclusions

A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.
Conclusions

A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.

reduce number of comparisons
Conclusions

A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.

reduce number of comparisons
Conclusions

A **proper metric** to enable **efficient similarity search** of process models based on a **behavioral abstraction** thereof.

**Definition 10 (Behavioral Profile Metric):** \( \mathcal{B} = (\mathcal{B}, d_B) \) is a metric space of behavioral profiles \( \mathcal{B} \), where the behavioral profile metric \( d_B : \mathcal{B} \times \mathcal{B} \to \mathbb{R} \) is a metric,

\[
d_B(B, B') = 1 - \sum w_k \cdot \text{sim}_h(B, B')
\]

with \( h \in \{ +, -, \ldots \} \) and weighting factors \( w_k \in \mathbb{R}, 0 < w_k < 1 \) such that \( \sum w_k = 1 \).

reduce number of comparisons

overlap in behavioral profile relations

lundi 5 septembre 2011
Conclusions

A proper metric to enable efficient similarity search of process models based on a behavioral abstraction thereof.

- Provide good quality and performance
- Reduce number of comparisons
- A suitable distance function
References

Fig. 1. Metric space partition $T(p)$ (solid circle) with pivot $p$ and similarity query (dotted circle) with query model $q$. (a) Exclusion, (b) Inclusion, (c) Intersection.
### Table 1. Overview of process model similarities

<table>
<thead>
<tr>
<th>Approach</th>
<th>Aspect</th>
<th>Symmetry</th>
<th>Nonnegativity</th>
<th>Identity</th>
<th>Triangle inequality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minor et al. [24]</td>
<td>Structure</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Li et al. [18]</td>
<td>Structure</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Ehrig et al. [25]</td>
<td>Structure</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Dijkman et al. [17]</td>
<td>Structure</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Eshuis and Grefen [22]</td>
<td>Behavior</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Aalst et al. [26]</td>
<td>Behavior</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Wombacher and Rozie [20]</td>
<td>Behavior</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Lu and Sadiq [27]</td>
<td>Structure</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Dongen et al. [23]</td>
<td>Behavior</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Nejati et al. [21]</td>
<td>Behavior</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Yan et al. [19]</td>
<td>Structure</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>
Process Model Similarity


Behavioral Profile (formally)

**Definition 3 (Weak Order).** Let \( P = (A, s, e, C, N, F, T) \) be a process model and \( T_P \) its set of traces. The weak order relation \( \succ_P \subseteq (A \times A) \) contains all pairs \( (x, y) \), such that there exists a trace \( \sigma = \langle a_1, a_2, \ldots \rangle \) in \( T_P \) and there exist two indices \( j, k \in \{1, 2, \ldots \} \) with \( j < k \) for which holds \( a_j = x \) and \( a_k = y \).

Using the notion of weak order, we define the behavioral profile as follows.

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

\( \mathcal{B}_P = \{\prec_P, +_P, ||_P\} \) is the behavioral profile of \( P \).
We define the behavioral profile matrix as the strict order relation $\mathcal{B}$ between activities $X$ and $Y$. This can be seen in the behavioral profile matrix as the strict order $\mathcal{B}(a, b) = 1$ for activities $a$ and $b$.

Definition 6: Let $P$, $Q$ be process models that feature the same exclusiveness relation $\mathcal{E}$. Then, the exclusiveness similarity $\sim$ is defined as

$$\sim(P, Q) = \frac{|P \cap Q|}{|P \cup Q|},$$

where $|P \cap Q|$ is the number of common activities in both process models $P$ and $Q$, and $|P \cup Q|$ is the total number of activities in both models.

Fig. 2. Order management process models $m_1$ and $m_2$ with their behavioral profile matrices $\mathcal{B}_{m_1}$ and $\mathcal{B}_{m_2}$.
Behavioral Profile Metric (formally)

**Definition 1 (Metric).** A metric is a distance function \( d: \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R} \) between objects of domain \( \mathcal{D} \) with the following properties:

- **Symmetry:** \( \forall o_i, o_j \in \mathcal{D} : d(o_i, o_j) = d(o_j, o_i) \)
- **Nonnegativity:** \( \forall o_i, o_j \in \mathcal{D} : o_i \neq o_j : d(o_i, o_j) > 0 \)
- **Identity:** \( \forall o_i, o_j \in \mathcal{D} : d(o_i, o_j) = 0 \iff o_i = o_j \)
- **Triangle inequality:** \( \forall o_i, o_j, o_k \in \mathcal{D} : d(o_i, o_k) \leq d(o_i, o_j) + d(o_j, o_k) \)

A metric space is a pair \( \mathcal{S} = (\mathcal{D}, d) \).

**Definition 10 (Behavioral Profile Metric).** \( \mathcal{B} = (\mathcal{B}, d_{\mathcal{B}}) \) is a metric space of behavioral profiles \( \mathcal{B} \), where the behavioral profile metric \( d_{\mathcal{B}}: \mathcal{B} \times \mathcal{B} \rightarrow \mathbb{R} \) is a metric,

\[
d_{\mathcal{B}}(\mathcal{B}_P, \mathcal{B}_Q) = 1 - \sum_h w_h \cdot \text{sim}_h(\mathcal{B}_P, \mathcal{B}_Q)
\]

with \( h \in \{+,-,||,\sim',||'\} \) and weighting factors \( w_h \in \mathbb{R}, 0 < w_h < 1 \) such that \( \sum_h w_h = 1 \).
Fig. 3. (a) Precision-recall curve for metrics based on elementary similarities, (b) precision-recall curve for a baseline metric, and two aggregated metrics.