Slice, Mine, Dice: Complexity-Aware Automated Discovery of Business Process Models

Chathura C. Ekanayake (QUT), Marlon Dumas (UT), Luciano García-Bañuelos (UT), Marcello La Rosa (QUT)
Discovering “all-in-one” process models...
Trace clustering

Process variant 1

Cluster 1
- a b a b c d e c a f g h
- a b a b k d e c h f g h

Cluster 2
- b c p q p r a k q r s
- b c p h p r a k q r s

Noise
- a x p h y z t t u

high complexity & redundancy
Slice, Mine and Dice (SMD)

1) Slice
2) Mine
3) Dice

Event log

Trace clustering

Complexity threshold e.g. Size ≤ 30

Process model

Slice, Mine and Dice (SMD)
Slice, Mine and Dice (SMD)

Event log

Trace clustering

Complexity threshold e.g. Size ≤ 30

Process models

Process discovery
Slice, Mine and Dice (SMD)

Event log

Trace clustering

Complexity threshold

e.g. Size ≤ 30

Process models

Process discovery

\( \leq \)
A closer look…

Process model M3

Process model M2
Refined Process Structure Tree (RPST)

Extracting exact clones

Exact clones:
- SESE
- Non-trivial
- Identical
Extracting approximate clones

Appr. clones:
- SESE
- Non-trivial
- Similar
- Unrelated

Merging approximate clones

The result...
Evaluation setup

Trace clustering

• M. Song, C.W. Gunther, and W.M.P. van der Aalst, Improving Process Mining with Trace Clustering, J. Korean Inst. of Industrial Engineers 34(4), 2008
• R.P.J.C. Bose, W.M.P. van der Aalst, Trace Clustering Based on Conserved Patterns: Towards Achieving Better Process Models, BPM 2009 Workshops

Discovery


<table>
<thead>
<tr>
<th>Log</th>
<th>Traces</th>
<th>Events</th>
<th>Event classes</th>
<th>Duplication ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motor</td>
<td>4,293</td>
<td>33,202</td>
<td>292</td>
<td>114</td>
</tr>
<tr>
<td>Commercial</td>
<td>4,852</td>
<td>54,134</td>
<td>81</td>
<td>668</td>
</tr>
<tr>
<td>BPI 2012</td>
<td>5,312</td>
<td>91,949</td>
<td>36</td>
<td>2,554</td>
</tr>
</tbody>
</table>
Evaluation – repository size and models number

• up to 64% reduction in repository size
• up to 66% reduction in # of top level process models
• up to 120 sub-processes extracted

S: Song et al.
B: Bose et al.
M: Medeiros et al.
### Evaluation – individual model complexity

<table>
<thead>
<tr>
<th>Log</th>
<th>Method</th>
<th>Size</th>
<th>CFC</th>
<th>ACD</th>
<th>Density</th>
<th>CNC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Avg</td>
<td>Min</td>
<td>Max</td>
<td>Savings (%)</td>
<td></td>
</tr>
<tr>
<td>Motor</td>
<td>S</td>
<td>22.75</td>
<td>4</td>
<td>37</td>
<td>22.8</td>
<td>12.07</td>
</tr>
<tr>
<td></td>
<td>S + SMD</td>
<td>17.57</td>
<td>4</td>
<td>37</td>
<td>10.07</td>
<td>2.34</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>20.01</td>
<td>4</td>
<td>37</td>
<td>9.97</td>
<td>2.51</td>
</tr>
<tr>
<td></td>
<td>B + SMD</td>
<td>18.04</td>
<td>4</td>
<td>37</td>
<td>10.05</td>
<td>2.38</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>15.73</td>
<td>3</td>
<td>49</td>
<td>-1.1</td>
<td>7.36</td>
</tr>
<tr>
<td></td>
<td>M + SMD</td>
<td>15.9</td>
<td>4</td>
<td>49</td>
<td>8.34</td>
<td>2.12</td>
</tr>
<tr>
<td>Commercial</td>
<td>S</td>
<td>24.07</td>
<td>6</td>
<td>34</td>
<td>22.4</td>
<td>13.65</td>
</tr>
<tr>
<td></td>
<td>S + SMD</td>
<td>18.67</td>
<td>2</td>
<td>34</td>
<td>11.34</td>
<td>2.49</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>21.11</td>
<td>2</td>
<td>34</td>
<td>20.3</td>
<td>11.04</td>
</tr>
<tr>
<td></td>
<td>B+ SMD</td>
<td>16.82</td>
<td>2</td>
<td>34</td>
<td>9.73</td>
<td>2.29</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>18.86</td>
<td>2</td>
<td>40</td>
<td>11.1</td>
<td>10.18</td>
</tr>
<tr>
<td></td>
<td>M + SMD</td>
<td>16.76</td>
<td>2</td>
<td>34</td>
<td>9.71</td>
<td>2.38</td>
</tr>
<tr>
<td>BPI</td>
<td>S</td>
<td>47.32</td>
<td>15</td>
<td>56</td>
<td>29.7</td>
<td>20.77</td>
</tr>
<tr>
<td></td>
<td>S + SMD</td>
<td>33.27</td>
<td>4</td>
<td>56</td>
<td>20.18</td>
<td>2.41</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>46.54</td>
<td>13</td>
<td>56</td>
<td>30.6</td>
<td>20.48</td>
</tr>
<tr>
<td></td>
<td>B + SMD</td>
<td>32.3</td>
<td>4</td>
<td>56</td>
<td>19.29</td>
<td>2.33</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>46.48</td>
<td>21</td>
<td>61</td>
<td>18.9</td>
<td>21.16</td>
</tr>
<tr>
<td></td>
<td>M + SMD</td>
<td>37.71</td>
<td>7</td>
<td>56</td>
<td>25.29</td>
<td>2.38</td>
</tr>
</tbody>
</table>
The SMD technique

+ Seeks to discover process models that meet user-specified complexity thresholds

+ Reduces repository size and top level models number compared to trace clustering techniques

+ Preserves fitness, appropriateness & generalization of models discovered from trace clusters

+ Little impact on structural complexity of process models

- Performance overhead
In the future...

Optimizations

- Parallelization of divisive trace clustering, discovery and GED computation
- Incremental divisive trace clustering
- Sub-polygons and sub-bonds extraction

Complexity thresholds tuning
Slice, Mine, Dice: Complexity-Aware Automated Discovery of Business Process Models
SMD complexity

The complexity of SMD depends on:

- Process discovery – \( c_1 \)
- Exact clone detection – \( c_2 \)
- Approximate clone detection – \( c_3 \)
- Process merging – \( c_4 \)

Let:

- \(|D| = \text{number of trace clusters in the dendrogram}\)
- \(|F| = \text{number of fragments in all process models discovered from } D\)

Worst-case complexity = \( O(|D|c_1 + |F|(c_2 + c_4) + |F|^2c_3) \)
# Related work

<table>
<thead>
<tr>
<th>Method</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trace clustering</strong></td>
<td>• Redundancies between variants are not identified</td>
</tr>
<tr>
<td>(Song et al., Bose et al.)</td>
<td>• May produce unnecessarily complex collection</td>
</tr>
<tr>
<td></td>
<td>• Used in the initial stage of our technique</td>
</tr>
<tr>
<td><strong>Log level abstractions</strong></td>
<td>• Identifies similarities in an event log</td>
</tr>
<tr>
<td>(Bose et al.)</td>
<td>• Does not control complexity</td>
</tr>
<tr>
<td><strong>Hierarchy abstractions</strong></td>
<td>• Focuses on building a navigable hierarchy</td>
</tr>
<tr>
<td>(Greco el al.)</td>
<td>• Does not perform refactoring, only abstracts if commonalities exist</td>
</tr>
<tr>
<td></td>
<td>• Does not control complexity</td>
</tr>
<tr>
<td><strong>Fuzzy miner</strong></td>
<td>• Process elements are grouped based on frequency</td>
</tr>
<tr>
<td>(Günther et al.)</td>
<td>• Does not control complexity</td>
</tr>
<tr>
<td></td>
<td>• Does not create concrete process models</td>
</tr>
</tbody>
</table>
Finding approximate clones

Any fragment in a cluster should not contain any other fragment in the same cluster

DBSCAN - Extension
Distance measure - Graph Edit Distance (GED)

Step 1: Remove edge
Step 2: Add task
Step 3: Add edge
Step 4: Add edge

Absolute distance = 4
Normalized distance = 4 / (10 + 12) = 0.18
Evaluation – Motor log

- 14+% reduction in repository size
- 20+% reduction in # of top level process models
- up to 104 sub-processes extracted

S: Song et al.  
B: Bose et al.  
M: Medeiros et al.
Evaluation – Commercial log

- 19+% reduction in repository size
- 22+% reduction in top level process models
- up to 121 sub-processes extracted

S: Song et al.
B: Bose et al.
M: Medeiros et al.
Evaluation – BPI 2012 log

- 60+% reduction in repository size
- 63+% reduction in top level process models
- up to 51 sub-processes extracted

S: Song et al.
B: Bose et al.
M: Medeiros et al.