Flexible Evolutionary Algorithms for Mining Structured Process Models

Joos Buijs

With Boudewijn van Dongen and Wil van der Aalst
431 403 municipalities
Germany: +/-12,000
U.S.: +/-20,000

Software as a Service (SaaS)
Traditional Situation

Processes
Municipality 1
IS1
E1
M1

Processes
Municipality 2
IS2
E2
M2

Processes
Municipality n
ISn
En
Mn
Our Vision
7 Challenges

Processes Municipality 1

Processes Municipality 2

Processes Municipality n

Process Discovery

Compare with Documented Process Model

Describe Family of Processes
### Process Discovery

#### Trace #

<table>
<thead>
<tr>
<th>Trace</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E G</td>
<td>60</td>
</tr>
<tr>
<td>A B D C E G</td>
<td>120</td>
</tr>
<tr>
<td>A B C F G</td>
<td>80</td>
</tr>
<tr>
<td>A D B C F G</td>
<td>10</td>
</tr>
<tr>
<td>A D C B F G</td>
<td>40</td>
</tr>
<tr>
<td>A C B F G</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trace</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D F G</td>
<td>380</td>
</tr>
<tr>
<td>A B D C F G</td>
<td>260</td>
</tr>
<tr>
<td>A C B E G</td>
<td>10</td>
</tr>
<tr>
<td>A D B C E G</td>
<td>10</td>
</tr>
<tr>
<td>A C D B F G</td>
<td>20</td>
</tr>
<tr>
<td>A B C G</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trace</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>A C B D G</td>
<td>4</td>
</tr>
<tr>
<td>A D E G</td>
<td>4</td>
</tr>
<tr>
<td>A C F G</td>
<td>3</td>
</tr>
<tr>
<td>A B C D E F G</td>
<td>2</td>
</tr>
<tr>
<td>A B D E G</td>
<td>2</td>
</tr>
<tr>
<td>A C D F G</td>
<td>2</td>
</tr>
</tbody>
</table>

---

**Diagram:**

- **Nodes:** a, b, c, d, e, f, g
- **Links:** a → b, b → c, c → d, d → e, e → f, f → g, g → a

**Notations:**
- '+' nodes represent sequential steps.
- 'X' nodes represent decision points.
Results on Real Data
Challenge 1: Discovering Correct Process Models

Heuristics Miner Result

Genetic Miner Result
Solution 1: Process Trees

- Always sound because of the block structure
- Also Loop and OR operator

A (BA)* C
Challenge 2:
Improve Representational Bias
Solution 2: Translate Process Trees to...

- Inherently structured
- Translation is straight-forward
Challenge 3: Balancing the Quality of Discovered Process Models

- "able to replay event log"
- "Occam’s razor"
- replay fitness
- simplicity
- generalization
- precision
- "not overfitting the log"
- "not underfitting the log"
Balancing Forces

Replay Fitness

Generalization

Simplicity

Precision
Balancing Forces

Replay Fitness

Simplicity

Generalization
The Evolutionary Tree Miner Framework

Create Initial Population

Change Population

Measure Quality

Stop?

Yes

No

Return Best Individual

Replay Fitness

Simplicity

Generalization

Precision
Challenge 4: Improve Understandability for Non-Experts

- Understandable process models (Challenge 2)
- Trustworthiness of result

Versus

(on process tree interpretation)
Solution 4a: Relate Event Log, Process Model and System

- Modeled System Behavior
- Non-Modeled System Behavior
- Modeled Non-System Behavior

- Precision
- Replay Fitness
- Noise

- 'Imprecise' Behavior
- Modeled and Observed Behavior
- 'Non-Replayable' Behavior

- Observed System Behavior
- Non-Observed System Behavior

- Generalization

- Model
- Event Log
- System
Solution 4b: Provide Insights in the Trade-Offs between the 4 Quality Dimensions
Solution 4b: Pareto Front on Running Example Event Log
ETM\textsubscript{d} on Real Data

Sound

Readable

One of 33 Models

\begin{tabular}{|c|c|}
\hline
F: & 0.881 \\
\hline
P: & 0.955 \\
\hline
S: & 1.000 \\
G: & 0.920 \\
\hline
\end{tabular}
7 Challenges

- Process Discovery
- ETMd
- Compare with Documented Process Model
- Describe Family of Processes

**IS1**
- Processes Municipality 1
  - E1
- IS2
  - Processes Municipality 2
  - E2
  - M2
- ISn
  - Processes Municipality n
  - En
  - Mn

- Processes Municipality 1
  - C1
- Processes Municipality 2
  - C2
- Processes Municipality n
  - Cn

Page 21
Challenge 5: Use Existing Knowledge in Process Discovery (ETMr)

Documented Process Model

Repaired Process Model (one of many)

ETMr : ETMd with similarity quality dimension

Observed Process
7 Challenges
Challenge 6: Describe a Family of Processes

- IS1
  - E1
  - M1
  - Processes Municipality 1

- IS2
  - E2
  - M2
  - Processes Municipality 2

- ISn
  - En
  - Mn
  - Processes Municipality n

...
Solution 6: Discover a Configurable Process Model (ETMc)

ETMc:
ETMd with
- Configurable process tree
- Configuration quality dimension
- Configuration change operations
Challenge 7: Compare Similar Observed Behavior

Replay behavior of Event Log / on Model $m$

With Hajo Reijers
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comparison Table

With Hajo Reijers
## Solution 7: Results on Running Example Variants

<table>
<thead>
<tr>
<th>Event log 1</th>
<th>Variant 1</th>
<th>Variant 2</th>
<th>Variant 3</th>
<th>Variant 4</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Event log 2</th>
<th>Variant 1</th>
<th>Variant 2</th>
<th>Variant 3</th>
<th>Variant 4</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Event log 3</th>
<th>Variant 1</th>
<th>Variant 2</th>
<th>Variant 3</th>
<th>Variant 4</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Event log 4</th>
<th>Variant 1</th>
<th>Variant 2</th>
<th>Variant 3</th>
<th>Variant 4</th>
</tr>
</thead>
</table>
7 Challenges

- Processes Municipality 1
- Processes Municipality 2
- Processes Municipality n

Process Discovery

Compare with Documented Process Model

Describe Family of Processes
7 Solutions

1. Discover Correct Process Models:
   √ Process Trees
2. Improve the Representational Bias
   √ Translate Process Trees
3. Balancing the Quality of Discovered Process Models
   √ ETMd
4. Improve Understandability for Non-Experts
   √ Discussion of the 4 Quality Dimensions
   √ Discovery of a Pareto Front
5. Use Existing Knowledge in Process Discovery
   √ ETMr
6. Describe a Family of Processes
   √ ETMc
7. Compare Similar Observed Behavior
   √ Comparison Framework
Future Work

• Increase expressiveness of Process Trees (add deferred choice, milestone, …, operators)
• Improve ETM performance
  • Smart(er) change of Process Trees
  • Alignment estimation
• Further philosophical discussion on generalization
• New ETM algorithms (data, simulation, …)
Outlook

- Academics need real data and real issues from tool vendors, consultants and industry
- Address hindering side issues (privacy etc.)
- Improve applicability and usability of academic solutions to real data by real people
- Incentives for academics to publish (quality) implementation and documentation
- Incorporate additional data/knowledge (might make certain things easier)
- Structured approach to discover a Pareto front