Examining Case Management Demand using Event Log Complexity Metrics

@Adaptive CM 2014 Workshop

Marian Benner-Wickner, Matthias Book, Tobias Brückmann, Volker Gruhn
Agenda

- Problem domain
- Event log complexity metrics
- Experimental evaluation
### Problem domain

- Industrialization of structured business processes
  - State-of-the-Art: Automation and workflow support for well-known processes

<table>
<thead>
<tr>
<th></th>
<th>Unknown</th>
<th>Well-known</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatable</td>
<td>Mine process, then ...</td>
<td>Automate!</td>
</tr>
<tr>
<td>Non-automatable</td>
<td>Mine process, then ...</td>
<td>Manage workflow!</td>
</tr>
</tbody>
</table>

- But what if...
  - ...mining fails, e.g. producing “spaghetti processes”?
  - ...actors permanently break/avoid workflow?
- Answer: Case Management (CM) techniques, e.g. Agenda-driven CM
How to spot the right support technology

- Today: The “try-and-error” way

```
Model available?
   yes
   no

Process Mining

Mining successful?
   yes
   no

Auto-matable?
   yes
   no

Automation (e.g. via BPEL)  Workflow Management  Case Management
```
How to spot the right support technology (2)

- Suggestion of a more reasonable, convenient way:
How to measure the demand for CM techniques using event logs?

Step 1: Define CM process characteristics

Step 2: Introduce event log metrics!

Step 3: Experimental evaluation
CM process characteristics

- Assumption: CM processes produce complex event logs
- Justification
  - Cases differ widely (diversity)
    - Each case produces a trace (=course of events) very different to all other traces
  - Information theory lingo: a CM process event log has a high entropy
- Two dimensions of diversity
  - Event Diversity: Within any given trace, there are many different events
  - Trace Diversity: Comparing any two traces, the course of events is very different (very few re-occurring pattern)
Event log complexity metrics

Event diversity

- **Definition 1.1 (Average trace length)**
  - \( k \): number of traces within the event log \( L \); \( k > 0 \).
  - \( n_i \): number of each trace’s events

\[
\text{atl}(L) = \frac{1}{k} \sum_{i=1}^{k} n_i.
\]

- Rehabilitation management example:
  - Complicated case (occupational accident)
  - Many events during various different therapies
  - Some events are re-occurring

- \( \text{atl}(L) \) does not reflect the structure
  - Metric biased by many reoccurrences due to process loops
Event log complexity metrics

Event diversity

- **Definition 1.2 (Average trace size)**
  - $k$: number of traces within the event log $L$
  - $E_i$: set of each trace’s distinct events

\[
ats(L) = \frac{1}{k} \sum_{i=1}^{k} |E_i|
\]

- “Mean number of distinct events”
- Benefit: robust against reoccurrence bias
Event log complexity metrics

Event diversity

• But what about the amount of unique events?
  ➔ Measure ratio between distinct events and total events

• **Definition 1.3 (Event density)**
  
  \[ ed(L) = \frac{ats(L)}{atl(L)} \]

• Consider both extremes:
  • Case A: every event is unique
    \[ ats(L) = atl(L) \Rightarrow ed(L) = 1 \ (high \ density) \]
  • Case B: Very few events which reoccur very often
    \[ ats(L) << atl(L) \Rightarrow ed(L) = 0 \ (low \ density) \]
Event log complexity metrics

Trace diversity

• How to quantify differences between traces of a log?
• Idea: Measure probability that a given event occurs in any trace

• Definition 2.1 (Simple trace diversity).
  • $n$: total number of distinct events
  
  $$\text{std}(L) = 1 - \frac{\text{ats}(L)}{n}.$$  

• Consider both extremes:
  • A few distinct events occur within each trace
    $$\text{ats}(L) \ll n \Rightarrow \text{std}(L) = 1 \text{ (high diversity)}$$
  • Every distinct event occurs in every trace (at least once)
    $$\text{ats}(L) = n \Rightarrow \text{std}(L) = 0 \text{ (low diversity)}$$
Event log complexity metrics

Trace diversity

- Yet, no metric considers the order of occurrence!
- Idea: Measure the mutual dissimilarity between all traces
- Approach:
  1. Treat traces as words and events as their characters
  2. Use levenshtein distance to measure dissimilarity

**Definition 2.2 (Advanced trace diversity).** Let $k$ be the total number of traces and $LD(T_i, T_j)$ be the levenshtein distance between two traces and $T_i$, $T_j$. Recall that $atl(L)$ is the average trace length of $L$. The advanced trace diversity is defined as follows:

$$
atd(L) = \frac{2}{k \times (k-1) \times atl(L)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} LD(T_i, T_j)$$
Event log complexity metrics

Trace diversity

• How to calculate LD(T_i, T_j):
  1. Replace event names by a dedicated unicode character for each name
  2. Calculate levenshtein distance, treating traces as words

• Example log L_1 = \{T_1, T_2, T_3, T_4, T_5\} =
  \{<a, b, c, d, e>, <f, g, h, i, j>, <k, l, m, n, o>, <p, q, r, s, t>, <u, v, w, x, y>\}

  k = 5, \text{atd}(L_1) = 5, each LD(T_i, T_j) value is 5

  \[\text{atd}(L_1) = \frac{2}{5 \times (5-1) \times 5} \sum_{i=1}^{5} \sum_{j=i+1}^{5} 5 = \frac{2}{100} \times 50 = 1\]

• Example log L_2 = \{T_1, T_2, T_3, T_4, T_5\} =
  \{<a, b, c, d, e>, <a, b, c, d, e>, <a, b, c, d, e>, <a, b, c, d, e>, <a, b, c, d, e>\}

  \[\text{atd}(L_2) = \frac{2}{5 \times (5-1) \times 5} \sum_{i=1}^{5} \sum_{j=i+1}^{5} 0 = \frac{2}{100} \times 0 = 0\]
Experimental evaluation

Activity

Generate artificial processes

Experiment with artificial processes

Collect real-life logs

Experiment with real-life logs

Output

Metrics are valid!

Metrics can indeed measure complexity in the field
# Experimental evaluation

## Detailed results

- **Detailed results (artificial logs):**

<table>
<thead>
<tr>
<th>Log</th>
<th>atl</th>
<th>ats</th>
<th>ed (%)</th>
<th>std (%)</th>
<th>atd (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L₁</td>
<td>13</td>
<td>13</td>
<td>100</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>L₂</td>
<td>13</td>
<td>7</td>
<td>54</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>L₃</td>
<td>6</td>
<td>6</td>
<td>100</td>
<td>84</td>
<td>15</td>
</tr>
<tr>
<td>L₄</td>
<td>11</td>
<td>8</td>
<td>73</td>
<td>60</td>
<td>18</td>
</tr>
</tbody>
</table>

- **Detailed results (real-life logs):**

<table>
<thead>
<tr>
<th>Log</th>
<th>atl</th>
<th>ats</th>
<th>ed (%)</th>
<th>std (%)</th>
<th>atd (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L₅</td>
<td>131</td>
<td>33</td>
<td>25</td>
<td>95</td>
<td>37</td>
</tr>
<tr>
<td>L₆</td>
<td>20</td>
<td>12</td>
<td>60</td>
<td>67</td>
<td>17</td>
</tr>
<tr>
<td>L₇</td>
<td>9</td>
<td>4</td>
<td>44</td>
<td>69</td>
<td>19</td>
</tr>
</tbody>
</table>

Ad-hoc, mining impossible
Structured, mining possible
Ad-hoc (but mining successful)
Experimental evaluation

Detailed results

- Metrics indicate demand for alternative techniques like CM, however:

  1. Keep the overall picture: Do not consider one metric individually!
  2. Careful: Event/trace diversity metrics correlate!

  Abstraction leads to false positives!
  Noise leads to false negatives!

Model available?
- yes
- no

Sufficient event log quality?
- yes
- no

Process Mining

Auto-\(\text{m}a\)tatable?
- yes
- no

Automation (e.g. via BPEL)
Workflow Management
Case Management
Thank you!

Questions?