Filtering Out Chaotic Activities From Event Logs

Niek Tax
Motivating Example

Event sequences

\(<X,A,B,D,E,H>\)
\(<A,D,C,E,G>\)
\(<A,X,C,D,E,F,B,D,X,E,G>\)
\(<A,D,B,E,H>\)
\(<A,C,D,E,F,D,C,E,F,C,D,E,X,H>\)
\(<A,C,X,D,E,G>\)

A) Register request
B) Examine casually
C) Example thoroughly
D) Check ticket
E) Decide
F) Re-initiate request
G) Pay compensation
H) Reject request

What if we would have logged one additional activity:

X) The customer calls

Generated from Inductive Miner
Outline

• Motivating Example

• Existing Event Log Filtering Techniques

• Chaotic Activity Filtering Techniques

• Evaluation on Synthetic Data

• Evaluation on Real Life Data

• Conclusions
Existing Work on Event Log Filtering

• Three categories:
  1. Event filtering
  2. Activity filtering
  3. Trace filtering
Frequency-based activity filtering doesn’t help: chaotic activities can be frequent (or infrequent)
Event filtering doesn't help: directly-follows relations between other activities are obfuscated
Trace Filtering

- A chaotic activity might be so frequent that it occurs in (almost) all traces, in which case trace filtering fundamentally cannot solve this issue.
Outline

- Motivating Example
- Existing Event Log Filtering Techniques
- Chaotic Activity Filtering Techniques
- Evaluation on Synthetic Data
- Evaluation on Real Life Data
- Conclusions
Basic Concepts

- **Subtrace frequency**
  \[ \#(\sigma, L) = \sum_{\sigma' \in L} \{0 \leq i \leq |\sigma'| - |\sigma| \bigg| \forall 1 \leq j \leq |\sigma| \sigma'(i+j) = \sigma(j)\} \]

- **Directly-follows ratio**
  \[ dfr(a, b, L) = \frac{\#(\langle a, b \rangle, L)}{\#(a, L)} \]

- **Directly-precedes ratio**
  \[ dpr(a, b, L) = \frac{\#(\langle b, a \rangle, L)}{\#(a, L)} \]

- **Artificial start and end**
  - For each trace \( \sigma = \langle e_1, e_2, \ldots, e_n \rangle \) in a log \( L \)
  - Log \( L \) contains a trace \( \sigma^\uparrow = \langle e_1, e_2, \ldots, e_n, \rangle \)
  - Log \( L \) contains a trace \( \sigma^\downarrow = \langle \rangle, e_1, \ldots, e_n \rangle \)
Directly-Follows Ratio Vector

Assuming an arbitrary but consistent ordering over \( b \in Activities(L) \cup \{ \} \)
\[ dfr(a, L) \] denotes a vector where each element has value \( dfr(a, b, L) \)

- **Example:**
  - Given log \( L = [\langle a, b, c, x \rangle^{10}, \langle a, b, x, c \rangle^{10}, \langle a, x, b, c \rangle^{10}] \)
  - Using arbitrary but consistent ordering \( \langle a, b, c, x, \rangle \)
  - \( dfr(a, L) = \langle 0, \frac{20}{30}, 0, \frac{10}{30}, 0 \rangle \)
  - 20 out of 30 \( a \) events are followed by \( b \)
  - 10 out of 30 \( a \) events are followed by \( b \)
Directly-Precedes Ratio Vector

Assuming an arbitrary but consistent ordering over $b \in \text{Activities}(L) \cup \{[]\}$, $dpr(a, L)$ denotes a vector where each element has value $dpr(a, b, L^\uparrow)$

- **Example:**
  - Given log $L = [\langle a, b, c, x \rangle^{10}, \langle a, b, x, c \rangle^{10}, \langle a, x, b, c \rangle^{10}]$
  - Using arbitrary but consistent ordering $\langle a, b, c, x, [] \rangle$
  - $dpr(a, L) = \langle 0, 0, 0, 0, 1 \rangle$
  - All $a$ events are preceded by artificial start
Information Entropy as a Measure of Chaoticness

- Entropy of categorical probability distribution
  \[ H(X) = -\sum_{x \in X} x \log_2(x) \]

- Given example log \( L = [\langle a, b, c, x \rangle^{10}, \langle a, b, x, c \rangle^{10}, \langle a, x, b, c \rangle^{10}] \)
  \[
  \text{dfr}(a, L) = \langle 0, \frac{20}{30}, 0, \frac{10}{30}, 0 \rangle \\
  \text{dpr}(a, L) = \langle 0, 0, 0, 0, 1 \rangle
  \]

- Entropy:
  \[
  H(\text{dfr}(a, L)) = 0.918 \\
  H(\text{dpr}(a, L)) = 0
  \]

- Activity chaoticness:
  \[
  H(a, L) = H(\text{dfr}(a, L)) + H(\text{dpr}(a, L))
  \]

<table>
<thead>
<tr>
<th>Activity</th>
<th>Chaoticness</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.918</td>
</tr>
<tr>
<td>b</td>
<td>1.837</td>
</tr>
<tr>
<td>c</td>
<td>1.837</td>
</tr>
<tr>
<td>x</td>
<td>3.170</td>
</tr>
</tbody>
</table>
Two Algorithms to Filter Activities based on Activity Chaoticness

• **Direct approach**

**Algorithm 1** An activity filtering approach based on entropy.

**Input:** event log $L$

**Output:** list of event logs $Q$

*Initialisation:*

1. $L' = L$
2. $Q = \langle L' \rangle$

*Main Procedure:*

3. **while** $|Activities(L')| > 2$ **do**
4. $acts = Activities(L')$
5. $a' = \arg \max_{a \in acts} H(a, L')$
6. $L' = L' \downarrow_{acts\setminus\{a'\}}$
7. $Q = Q \cdot \langle L' \rangle$
8. **end while**
9. **return** $Q$
Two Algorithms to Filter Activities based on Activity Chaoticness

- **Indirect approach**

**Algorithm 2** An indirect activity filtering approach based on entropy.

**Input:** event log $L$

**Output:** list of event logs $Q$

*Initialisation:*

1. $L’ = L$
2. $Q = \langle L’ \rangle$

*Main Procedure:*

3. **while** $|\text{Activities}(L’)| > 2$ **do**
4. $acts = \text{Activities}(L’)$
5. $a’ = \arg \min_{a \in acts} H(\langle L’ \ast acts \setminus \{a\} \rangle)$
6. $L’ = \langle L’ \ast acts \setminus \{a’\} \rangle$
7. $Q = Q \ast (\langle L’ \rangle)$
8. **end while**
9. **return** $Q$

with $H(L) = \sum_{a \in \text{Activities}(L)} H(a, L)$
Laplace Smoothing

- So far we have used the empirical estimate of the directly-follows and directly-precede ratios:
  \[ dfr(a, b, L) = \frac{\#(\langle a, b \rangle, L)}{\#(a, L)} \]

- Leads to an unreliable estimate when \( \#(a, L) \) is small

- Laplace estimator
  \[ dfr^s(a, b, L) = \frac{\alpha + \#(\langle a, b \rangle, L)}{\alpha (1 + |Activities(L)|) + \#(a, L)} \text{ with } \alpha \in \mathbb{R}_{\geq 0} \]

- From a Bayesian statistics point of view, this is equivalent to using a Dirichlet prior to estimate the categorical probability distribution over the activity of the next event
  (with alpha the concentration parameter of the Dirichlet distribution)

Assuming an arbitrary but consistent ordering over \( b \in Activities(L) \cup \{\} \)\)

- \( dfr^s(a, L) \) denotes a vector where each element has value \( dfr^s(a, b, L) \)

\[
H^s(a, L) = H(dfr^s(a, L)) + H(dpr^s(a, L))
\]
Outline

• Motivating Example

• Existing Event Log Filtering Techniques

• Chaotic Activity Filtering Techniques

• Evaluation on Synthetic Data

• Evaluation on Real Life Data

• Conclusions
Evaluation with Synthetic Data

- Maruster A12

- Maruster A22

\(L_{A12}\)

Inductive Miner

Log of 25 traces

\(L_{A22}\)

Inductive Miner

Log of 400 traces
Experimental Setup

Three step experiment:

1. Add $k$ artificial chaotic activities to the log
2. Apply chaotic activity filtering technique until all $k$ artificially added chaotic activities are removed again
3. Measure how many other activities we removed from the log as a side-effect

Repeat three step experiment for:

- $L_{A12}$ and $L_{A22}$
- Different values of $k$
- Different frequency distributions for the chaotic activities
  - Frequent chaotic activities
  - Infrequent chaotic activities
  - Frequency independent chaotic activities
## TABLE I

The number of incorrectly filtered activities per filtering approach on $L_{A12}$ and $L_{A22}$ with $k$ added Frequency-independent (FI) / Frequent (F) / Infrequent (I) chaotic activities.

<table>
<thead>
<tr>
<th>Approach</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FI</td>
<td>F</td>
<td>I</td>
<td>FI</td>
<td>F</td>
<td>I</td>
<td>FI</td>
<td>F</td>
</tr>
<tr>
<td>Direct</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Direct ($\alpha = \frac{1}{</td>
<td>A</td>
<td>}$)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Indirect</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Indirect ($\alpha = \frac{1}{</td>
<td>A</td>
<td>}$)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Least-frequent-first</td>
<td>3</td>
<td>12</td>
<td>0</td>
<td>11</td>
<td>12</td>
<td>0</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Most-frequent-first</td>
<td>11</td>
<td>0</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>12</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Approach</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FI</td>
<td>F</td>
<td>I</td>
<td>FI</td>
<td>F</td>
<td>I</td>
<td>FI</td>
<td>F</td>
</tr>
<tr>
<td>Direct</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Direct ($\alpha = \frac{1}{</td>
<td>A</td>
<td>}$)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Indirect</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Indirect ($\alpha = \frac{1}{</td>
<td>A</td>
<td>}$)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Least-frequent-first</td>
<td>16</td>
<td>22</td>
<td>0</td>
<td>17</td>
<td>22</td>
<td>0</td>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>Most-frequent-first</td>
<td>7</td>
<td>0</td>
<td>22</td>
<td>8</td>
<td>0</td>
<td>22</td>
<td>19</td>
<td>0</td>
</tr>
</tbody>
</table>
Evaluation without Ground Truth

Experimental Setup:
1. Filter activities from the log one-by-one
2. Discover a process model from the filtered log after each filtering step
3. Measure *fitness/precision/f-score* of the process model with respect to the filtered log

- *Fitness* and *precision* measure the quality of the process model with respect to the event log
- Two different process models discovered from two different logs are not comparable in terms of *fitness* and *precision*
The Nondeterminism Metric

- **Align** the log on the model
- For each *synchronous move* in the alignment:
  - Determine which other activities would also have been possible to fire instead:
    - Go back to the marking after the previous synchronous move, check which non-silent transitions are reachable

```
Trace  | A | F | B
Model  | A | >> | B
Options| 1 | 4 |
```

Nondeterminism = (1+4) / 2 = 2.5
A Ground-Truth Free Evaluation

Nondeterminism results of the Inductive Miner models discovered from the Maruster A12 logs

Minimum % of activities explained
Outline

• Motivating Example

• Existing Event Log Filtering Techniques

• Chaotic Activity Filtering Techniques

• Evaluation on Synthetic Data

• Evaluation on Real Life Data

• Conclusions
# Evaluation on Real Life Data

## TABLE II

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
<th># traces</th>
<th># events</th>
<th># activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPI’12 [23]</td>
<td>Business</td>
<td>13087</td>
<td>164506</td>
<td>23</td>
</tr>
<tr>
<td>BPI’12 resource 10939 [24]</td>
<td>Business</td>
<td>49</td>
<td>1682</td>
<td>14</td>
</tr>
<tr>
<td>Environmental permit [25]</td>
<td>Business</td>
<td>1434</td>
<td>8577</td>
<td>27</td>
</tr>
<tr>
<td>SEPSIS [26]</td>
<td>Business</td>
<td>1050</td>
<td>15214</td>
<td>16</td>
</tr>
<tr>
<td>Traffic Fine [27]</td>
<td>Business</td>
<td>150370</td>
<td>561470</td>
<td>11</td>
</tr>
<tr>
<td>Bruno [28]</td>
<td>Human behavior</td>
<td>57</td>
<td>553</td>
<td>14</td>
</tr>
<tr>
<td>CHAD 1600010 [29]</td>
<td>Human behavior</td>
<td>26</td>
<td>238</td>
<td>10</td>
</tr>
<tr>
<td>Ordonez A [31]</td>
<td>Human behavior</td>
<td>15</td>
<td>409</td>
<td>12</td>
</tr>
<tr>
<td>van Kasteren [32]</td>
<td>Human behavior</td>
<td>23</td>
<td>220</td>
<td>7</td>
</tr>
<tr>
<td>Cook hh102 labour [33]</td>
<td>Human behavior</td>
<td>18</td>
<td>576</td>
<td>18</td>
</tr>
<tr>
<td>Cook hh102 weekend [33]</td>
<td>Human behavior</td>
<td>18</td>
<td>210</td>
<td>18</td>
</tr>
<tr>
<td>Cook hh104 labour [33]</td>
<td>Human behavior</td>
<td>43</td>
<td>2100</td>
<td>19</td>
</tr>
<tr>
<td>Cook hh104 weekend [33]</td>
<td>Human behavior</td>
<td>18</td>
<td>864</td>
<td>19</td>
</tr>
<tr>
<td>Cook hh110 labour [33]</td>
<td>Human behavior</td>
<td>21</td>
<td>695</td>
<td>17</td>
</tr>
<tr>
<td>Cook hh110 weekend [33]</td>
<td>Human behavior</td>
<td>6</td>
<td>184</td>
<td>14</td>
</tr>
</tbody>
</table>
Business Logs

Filter
- Direct
- Direct (a=1/|A|)
- Indirect
- Indirect (a=1/|A|)
- Least-frequent-first
- Flower model

Minimum % of activities explained

Nondeterminism
Van Kasteren Results

- Inductive Miner model on the log filtered until 4 activities were remaining with the indirect entropy-based filter with Laplace smoothing.

<table>
<thead>
<tr>
<th>Order</th>
<th>Filtered activity (indirect entropy-based filter with Laplace smoothing)</th>
<th>Filtered activity (least-frequent-first filter)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Use toilet</td>
<td>Prepare dinner</td>
</tr>
<tr>
<td>2</td>
<td>Get drink</td>
<td>Get drink</td>
</tr>
<tr>
<td>3</td>
<td>Leave house</td>
<td>Prepare breakfast</td>
</tr>
<tr>
<td>4</td>
<td>Take shower</td>
<td>Take shower</td>
</tr>
<tr>
<td>5</td>
<td>Go to bed</td>
<td>Go to bed</td>
</tr>
<tr>
<td>6</td>
<td>Prepare breakfast</td>
<td>Leave house</td>
</tr>
<tr>
<td>7</td>
<td>Prepare dinner</td>
<td>Use toilet</td>
</tr>
</tbody>
</table>

- Inductive Miner model on the log filtered until 4 activities were remaining with filtering infrequent activities.
Meta-Analysis Methodology

• Winning Number
  - The number of other filtering methods that are outperformed by a given filtering method for a given event log and percentage of activities

• For each percentage of activities remained:
  - For each filtering method:
    - Determine the average winning number over all logs
Meta-Analysis Results

![Graph showing the average winning number against the minimum percentage of activities explained for different filters.]
Possible Use Case for Chaotic Activity Filters
Conclusions

• The standard way of activity filtering, removing infrequent activities, is not the best way when aiming for behaviorally constrained process models.

• We have introduced four activity filtering techniques that outperform filtering out infrequent activities.

• The Indirect Entropy-based activity filter on average over all event logs included in the experiments outperforms the other techniques.

• All four introduced activity filtering techniques are available in ProM: https://svn.win.tue.nl/repos/prom/Packages/ActivityFiltering/
Should I analyze this event log?

- Sometimes analysis of an event log does not lead to any insights.

- “Is there signal in the data and did I do a poor job analyzing the data? Or is the data just random noise from which no insights can be gained?”

- A solution from the Time Series Analysis domain:
  The auto-correlation plot
Event Data as Categorical Time Series

- Association measures for categorical variables:

  **Chi-square measure**
  - Not normalized, therefore, two chi-square values between different variables are not comparable

  **Cramer’s V**
  - A normalized version of the Chi-square measure

  **Goodman and Kruskal’s tau**
  - Intuitive interpretation: percentage of variability of variable Y that can be explained by variable X
  - Not symmetric

  **Uncertainty Coefficient**
  - Many of the same properties as Goodman and Kruskal’s tau
  - Known to systematically underestimate association
Auto-Association Plot

Auto-association plot

Association vs. Lag

Trace ratio
Goodman and Kruskal's tau
ProM visualizer

- The visualizer “Visualize Log as an Auto-Association Plot” is now available in ProM:
  https://svn.win.tue.nl/repos/prom/Packages/AutoAssociationPlot/