Imperfection repair using dedicated and vanilla operators

We demonstrate the difference between event log repair using dedicated operators and using only 'vanilla' relational algebra operators. For three common patterns of imperfection in event logs, we show how each can be repaired with dedicated operators and how they can be repaired using only vanilla operators. Comparing the resulting statements shows the differences in suitability and computational complexity between the two sets of operators.

Homonymous Labels

Consider the example event log from Figure 1, which contains homonymous labels. Using the dedicated operators, we need the following statements to repair the event log:

$R_1 = \pi_{\uparrow cid,\uparrow act,\uparrow time,\uparrow res,\downarrow act}(<_{cid,time}(R_i))$

$R_2 = R_1 \cup \pi_{\uparrow cid,\uparrow act,\uparrow time,\uparrow res}(R_i \setminus (\pi_{\uparrow cid,\uparrow act,\uparrow time,\uparrow res}(R_1)))$

$R_o = \pi_{cid:\uparrow cid,act:f(\uparrow act,\uparrow res,\downarrow act),time:\uparrow time,res:\uparrow res}(R_2)$, where with slight abuse of notation:

$$
\begin{aligned}
f = \{&(\textit{'Review estimate'}, x, y) \mapsto \textit{'Manual review estimate'} \mid \\
&x \in Dom(\downarrow res) \setminus \{\textit{'System'}\} \wedge y \in Dom(\uparrow act) \\
\} \cup \\
\{&(\textit{'Review estimate'}, \textit{'System'}, y) \mapsto \textit{'Periodic review estimate'} \mid \\
&y \in Dom(\uparrow act) \setminus \{\textit{'Document received'}\} \\
\} \cup \\
\{&(\textit{'Review estimate'}, \textit{'System'}, \textit{'Document received'}) \mapsto \\
&\textit{'Automatic review estimate'} \\
\}
\end{aligned}
$$

cid	act	time	res
1	Document received	2202-01-31 11:30:45	100
1	Review estimate	2202-01-31 11:31:10	System
2	Document received	2202-01-31 13:15:30	100
2	Review estimate	2022-01-31 13:15:40	System
1	Review completed	2022-02-01 08:00:10	100
2	Review completed	2022-02-01 08:00:10	100
3	Document received	2202-04-11 08:30:45	100
3	Review estimate	2202-04-11 08:31:05	System
3	Review completed	2022-04-12 08:00:10	100
4	Review estimate	2202-04-12 09:15:30	200
4	Review completed	2202-04-19 11:25:30	300
1	Review estimate	2202-07-01 00:10:00	System
2	Review estimate	2202-07-01 00:10:01	System
3	Review estimate	2202-07-01 00:10:02	System

Figure 1: The homonymous label *'Review estimate'* in an insurance log (R_i).

In contrast, using only vanilla relational algebra operators, we need the following sequence of statements to repair the event log.

$$
\begin{aligned}
R_1 &= \pi_{cid,\uparrow act:act,\uparrow time:time,\uparrow res:res}(R_i)\\
R_2 &= \pi_{cid,\downarrow act:act,\downarrow time:time,\downarrow res:res}(R_i)\\
R_3 &= \pi_{cid,\uparrow act,\uparrow time,\uparrow res,\downarrow act,\downarrow time}(\sigma_{\uparrow time>\downarrow time}(R_1 \bowtie R_2))\\
R_4 &= R_3 \setminus (\pi_{Schema(R_3)}(\sigma_{time<\uparrow time \wedge time>\downarrow time}(R_3 \bowtie R_i)))\\
R_5 &= \pi_{cid,act:\uparrow act,time:\uparrow time,res:\uparrow res,\downarrow act}(R_4)\\
R_6 &= R_5 \cup \pi_{cid,act,time,res,\text{“”}}(R_i \setminus (\pi_{cid,act,time,res}(R_5)))\\
R_o &= \pi_{cid,act,time,res}(\sigma_{act\neq \text{`Review estimate'}}(R_6)) \cup\\
&\quad \pi_{cid,act:\text{`Manual review estimate'},time,res}(\\
&\qquad \sigma_{act=\text{`Review estimate'} \wedge res\neq \text{`System'}}(R_6)\\
&\quad) \cup\\
&\quad \pi_{cid,act:\text{`Periodic review estimate'},time,res}(\\
&\qquad \sigma_{act=\text{`Review estimate'} \wedge res=\text{`System'} \wedge \downarrow act\neq \text{`Document received'}}(R_6)\\
&\quad) \cup\\
&\quad \pi_{cid,act:\text{`Automatic review estimate'},time,res}(\\
&\qquad \sigma_{act=\text{`Review estimate'} \wedge res=\text{`System'} \wedge \downarrow act=\text{`Document received'}}(R_6)\\
&\quad)
\end{aligned}
$$

Collateral Events

Consider the example event log from Figure 2, which contains collateral events. Using the dedicated operators, we need the following statements to repair the event log:

$$
\begin{aligned}
R_o &= {}_{GC(t_1,t_2)}\mathcal{G}_{cid\mapsto any_{cid},act\mapsto anew,time\mapsto min_{time}} R_i, \text{ where}\\
GC(t_1,t_2) &= t_1[cid] = t_2[cid] \wedge\\
&\qquad\quad t_1[act] \in \mathcal{L} \wedge t_2[act] \in \mathcal{L} \wedge\\
&\qquad\quad |t_1[time] - t_2[time]| \leq \delta\\
\mathcal{L} &= \{\text{`Adjust recovery cost'}, \text{`Email'}, \text{`Pay assessor fee'}\}\\
\delta &= 600\\
anew(act) &= \text{`Pay insurance claim assessor'}
\end{aligned}
$$

cid	act	time
1	Submit claim	18/06/2014 03:20:09
1	Adjust recovery cost	19/06/2014 12:15:18
1	Email	19/06/2014 12:16:53
1	Pay assessor fee	19/06/2014 12:19:25
1	Approve claim	20/06/2014 8:40:01
2	Submit claim	20/06/2014 13:08:39
2	Email	21/06/2014 10:15:28
2	Adjust recovery cost	21/06/2014 10:16:43
2	Pay assessor fee	21/06/2014 10:19:53
2	Reject claim	22/06/2014 13:45:24

cid	act	time
1	Submit claim	18/06/2014 03:20:09
1	**Pay insurance claim assessor**	19/06/2014 12:15:18
1	Approve claim	20/06/2014 8:40:01
2	Submit claim	20/06/2014 13:08:39
2	**Pay insurance claim assessor**	21/06/2014 10:15:28
2	Reject claim	22/06/2014 13:45:24

Figure 2: Collateral Events, before (R_i) and after (R_o) repair using the generalised group-by operator.

In contrast, using only vanilla relational algebra operators, we need the following sequence of statements to repair the event log.

$$
\begin{aligned}
R_1 &= \pi_{cid, act_1:act, time_1:time}(\sigma_{act='Adjust\ recovery\ cost'}(R_i)) \\
R_2 &= \pi_{cid, act_2:act, time_2:time}(\sigma_{act='Email'}(R_i)) \\
R_3 &= \pi_{cid, act_3:act, time_3:time}(\sigma_{act='Pay\ assessor\ fee'}(R_i)) \\
R_4 &= \sigma_{|Max(time_1,time_2,time_3)-Min(time_1,time_2,time_3)|<\delta}(R_1 \bowtie R_2 \bowtie R_3) \\
R_5 &= \pi_{cid, act:'Pay\ insurance\ claim\ assessor', time:f_{time}(time_1,time_2,time_3)}(R_4) \\
R_6 &= \pi_{cid,act,time}(\sigma_{act=act_1 \wedge time=time_1}(R_i \bowtie R_4)) \cup \\
&\quad \pi_{cid,act,time}(\sigma_{act=act_2 \wedge time=time_2}(R_i \bowtie R_4)) \cup \\
&\quad \pi_{cid,act,time}(\sigma_{act=act_3 \wedge time=time_3}(R_i \bowtie R_4)) \\
R_o &= (R_i \setminus R_6) \cup R_5
\end{aligned}
$$

Form-based Event Capture

Consider the example event log from Figure 3, which contains form-based event capture. Using the dedicated operators, we need the following statements to repair the event log:

$$
\begin{aligned}
R_o = {}& \mathop{\mathcal{G}}_{GC(t_1,t_2)}^{\substack{cid \mapsto any_{cid}, act \mapsto anew, temp \mapsto temp, \\ pulse \mapsto pulse, bp \mapsto bp, time \mapsto min_{time}}} R_i, \text{ where} \\
& GC(t_1,t_2) = (t_1[cid] = t_2[cid]) \wedge \\
& \qquad t_1[act] \in \mathcal{L} \wedge t_2[act] \in \mathcal{L} \wedge t_1[time] = t_2[time] \\
& \mathcal{L} = \{\text{'Temperature'}, \text{'Pulse'}, \text{'Blood pressure'}\} \\
& anew(act) = \text{'Vital signs check'} \\
& temp(ps) = v \text{ if } (\text{'Temperature'}, v) \in ps \text{ else } \perp \\
& pulse \text{ and } bp \text{ are defined analogously.}
\end{aligned}
$$

cid	act	val	time
1	Triage		11/07/2015 04:20:09
1	Temperature	38	11/07/2015 06:12:18
1	Pulse	89	11/07/2015 06:12:18
1	Blood pressure	117	11/07/2015 06:12:18
1	Operation		11/07/2015 17:45:03
1	Temperature	37	11/07/2015 20:05:39
1	Pulse	81	11/07/2015 20:05:39
1	Discharge		12/07/2015 11:17:42

cid	act	temp	pulse	bp	time
1	Triage				11/07/2015 04:20:09
1	**Vital signs check**	38	89	117	11/07/2015 06:12:18
1	Operation				11/07/2015 17:45:03
1	**Vital signs check**	37	81		11/07/2015 20:05:39
1	Discharge				12/07/2015 11:17:42

Figure 3: Form-based Event Capture, before (R_i) and after (R_o) repair using the generalised group-by operator

In contrast, using only vanilla relational algebra operators, we need the following sequence of statements to repair the event log.

$$
\begin{aligned}
R_1 &= \pi_{cid, act_1:act, val_1:val, time_1:time}(\sigma_{act='Temprature'}(R_i)) \\
R_2 &= \pi_{cid, act_2:act, val_2:val, time_2:time}(\sigma_{act='Pulse'}(R_i)) \\
R_3 &= \pi_{cid, act_3:act, val_3:val, time_3:time}(\sigma_{act='Blood\ pressure'}(R_i)) \\
R_4 &= \sigma_{Max(time_1,time_2,time_3)=Min(time_1,time_2,time_3)}(R_1 \rtimes R_2 \rtimes R_3) \\
R_5 &= \pi_{\substack{cid, val:'', act:'Vital\ signs\ check', temp:val_1, pulse:val_2, \\ bp:val_3, time:any(time_1,time_2,time_3)}}(R_4) \\
R_6 &= \pi_{cid,act,val,time}(\sigma_{act=act_1 \wedge val=val_1 \wedge time=time_1}(R_i \bowtie R_4)) \cup \\
&\quad \pi_{cid,act,val,time}(\sigma_{act=act_2 \wedge val=val_2 \wedge time=time_2}(R_i \bowtie R_4)) \cup \\
&\quad \pi_{cid,act,val,time}(\sigma_{act=act_3 \wedge val=val_3 \wedge time=time_3}(R_i \bowtie R_4))
\end{aligned}
$$

$$R_o = (\pi_{cid,act,val,temp,pulse,bp,time}(R_i \setminus R_6)) \cup R_5$$