Finding Similar and Dissimilar Events for Preprocessing Logs and Improving Mining Results -
A Snapshot of Thesis and Tools

Xixi Lu, Dirk Fahland, Wil van der Aalst
Eindhoven University of Technology
Agenda

• (5 min) I. Process discovery and challenges
• (5 min) II. Research problem
• (10 min) III. Approach and applications (Recall)
• (20 min) IV. Demo and Discussions
Process Discovery

1. Process
2. Record
3. Log
4. Discover
5. Process model
Challenges in Flexible Environments

1. High variety of behavior
2. Deviating behavior
3. Duplicated tasks
1. High Variety of Behavior

Record:

A → B → C
C → S → A

Discover
2. Deviating behavior

Record

A → B → C
A → B → C → A
A → B → C → B

Discover

A → B → C
B → A → C
C → A → B
3. Duplicated Tasks

Record

V → A → B → C → A → S
V → C → B → A → S
V → C → A → B → S

Discover
Challenges

System unknown!

We don’t know what is optimal!

1. High variety of behavior
2. Deviating behavior
3. Duplicated tasks

1. Cluster?
2. Filter?
3. Yet Another Discovery?

Having dedicated algorithms?
Challenges

Log -> Discover -> Process model
Research Problem and Approach

- Map similar events
- Detect deviations
- Refine labels
- Cluster
- Discover
- Log
- Log with Mappings
- Log'
- Process model
Mapping Between Events

Trace 1

V → A → B → C → A → S

Trace 2

V → C → B → A → S

Mapping
Quantify Similarity of Mapped Events

... based on “structural context of events”
Quantify Similarity of Mapped Events

... based on “structural context of events”

(1) Differences in neighbors
Quantify Similarity of Mapped Events

... based on “structural context of events”

(1) Differences in neighbors
(2) Differences in structure

Distance = 4

Distance = 3

Cost = 2+1
Quantify Similarity of Mapped Events

... based on “structural context of events”
(1) Differences in neighbors
(2) Differences in structure

Cost = 2 + 1 + 1 + 1 + 0

Distance = 2

Distance = 1
Quantify Similarity of Mapped Events

... based on “structural context of events”
(1) Differences in neighbors
(2) Differences in structure

Cost = 2+3
Cost = 4+6
Cost = 4+4
Cost = 0+3
Cost so far = 6+10+8+5+3
Quantify Similarity of Mapped Events

... based on “structural context of events”
(1) Differences in neighbors
(2) Differences in structure
(3) #Dissimilar events

There is an algorithm to compute an optimal mapping:

Proposing Approach

1. Map similar events
2. Cluster
3. Detect deviations
4. Refine labels
5. Discover
6. Process model
Cluster Traces Using Mappings and Fusion
Cluster Traces Using Mappings and Fusion

Hierarchical Cluster of traces
Proposing Approach

1. Map similar events
2. Detect deviations
3. Refine labels
4. Cluster
5. Discover
6. Process model
Detect Deviation Using Mappings

**Threshold:** “Deviating event” if < T% of cases

Deviating event for T = 30%
Proposing Approach

1. Detect imprecise labels
2. Refine horizontally
3. Refine vertically

Log -> Map similar events -> Log with Mappings

Cluster -> Detect deviations -> Refine labels -> Log' -> Discover

Process model
Refine Labels Horizontally

Imprecise label candidates \{A, B, C\}

a) Normalize costs w.r.t. maximal cost seen in the log
Refine Labels Horizontally

Imprecise label candidates \{A, B, C\}

- **Trace 0**: V → A → B → C → S
  - Costs: 0.5 → 0.5 → 0.5

- **Trace 1**: V → A → B → C → A → S
  - Costs: 1 → 1 → 1

- **Trace 2**: V → C → B → A → S
  - Costs: 0.75 → 0.75 → 0.75

- **Trace 3**: V → C → A → B → S
  - Costs: 0.75

---

a) Normalize costs w.r.t. maximal cost seen in the log

b) Set variant threshold, say, 0.8

c) Remove edges if cost > variant threshold
Refine Labels Horizontally

1. Normalize costs w.r.t. maximal cost seen in the log
2. Set variant threshold, say, 0.8
3. Remove edges if cost > variant threshold
Refine Labels Vertically

I) Set unfolding threshold, say, 60%

II) Refine labels if freq > unfolding threshold

Is this A1 a 2nd iteration of a loop?
Refine Labels Vertically

I) Set unfolding threshold, say, 40%

II) Refine labels if freq > unfolding threshold

Or is this a different task?
Proposing Approach

Log → Map similar events → Log with Mappings

Cluster
Detect deviations
Refine labels

Log' → Discover → Process model
IV. Demo and Discussion

• Plugin “Log to Model Explorer”
Limitations?

• Implementation inefficient
• Do not know what is happening in the background
• Can not handle large loops and large parallel branches

• Clustering
  • Do not know why these clusters

• Filtering
  • Do not support finding missing events (yet)

• Refining labels
  • ...

The Bigger Picture?
Questions and Feedback?