Process Mining in the Context of Web Services

Prof.dr.ir. Wil van der Aalst
Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
w.m.p.v.d.aalst@tue.nl
Outline

• Web services monitoring
• Process Mining
• Running example
• Discovery
• Conformance checking
• Reality Check
• Conclusion

The work of many people!

Thanks to Ton Weijters, Boudewijn van Dongen, Ana Karla Alves de Medeiros, Anne Rozinat, Christian Günter, Eric Verbeek, Ronny Mans, Minseok Song, Laura Maruster, Huub de Beer, Peter van den Brand, Jan Mendling, Andriy Nikolov, Jianmin Wang, Lijie Wen, Irene Vanderfeesten, Mariska Netjes, Steffi Rinderle, Walid Gaaloul, Gianluigi Greco, Antonella Guzzo, etc. etc.
Web Services Monitoring
Setting: Services, composition, and choreography
Example: IBM’s WebSphere Process Server architecture

- Common event infrastructure (CEI)
Logging events

- local/global
- messages/activities

Services use BPEL or not, may have a model or not, are known or not, and may deviate from what is expected or not.
Overview Process Mining
Software systems are the mirror image of the “world”
Dual role of process models

“realistic models of systems make sense if they are an adequate reflection of reality”
Event logs are a reflection of reality

“world”

- business processes
- people
- machines
- components
- organizations

supports/controls

software system

records events, e.g., messages, transactions, etc.

“logs are everywhere and there will be more …”

event logs
Examples:
Process mining: Linking events to models

- "world"
  - business processes
  - people
  - machines
  - components
  - organizations

- models
  - analyzes

- verification

- software system
  - supports/controls
  - records events, e.g., messages, transactions, etc.

- event logs

- process/system model
  - discovery
  - conformance
Toy example to explain basic idea:

Reviewing of papers for IPA workshop 😊
Event log:
• processes
  – process instances
• events

Per event:
• activity name
• (event type)
• (originator)
• (timestamp)
• (data)
<ProcessInstance id="51" description="">
    - <AuditTrailEntry>
        <WorkflowModelElement>invite reviewers</WorkflowModelElement>
        <EventType>start</EventType>
        <Timestamp>2006-08-28T00:00:00.000+01:00</Timestamp>
        <Originator>Mike</Originator>
    </AuditTrailEntry>
    - <AuditTrailEntry>
        <WorkflowModelElement>invite reviewers</WorkflowModelElement>
        <EventType>complete</EventType>
        <Timestamp>2006-08-31T00:00:00.000+01:00</Timestamp>
        <Originator>Mike</Originator>
    </AuditTrailEntry>
    - <AuditTrailEntry>
        <Data>
            <Attribute name="result">reject</Attribute>
        </Data>
        <WorkflowModelElement>get review 3</WorkflowModelElement>
        <EventType>complete</EventType>
        <Timestamp>2006-09-02T00:00:00.000+01:00</Timestamp>
        <Originator>Mary</Originator>
    </AuditTrailEntry>
    - <AuditTrailEntry>
        <WorkflowModelElement>time-out 1</WorkflowModelElement>
        <EventType>complete</EventType>
        <Timestamp>2006-09-03T00:00:00.000+01:00</Timestamp>
        <Originator/>
    </AuditTrailEntry>
    - <AuditTrailEntry>
    - <AuditTrailEntry>
    - <AuditTrailEntry>
Discovery

models analyzes discovery records describes configures implements analyzes supports/controls conformance verifies maintains

“world” business processes people machines components organizations models analyzes

software system

records events, e.g., messages, transactions, etc.

process/system model
discovery

event logs

conformance
No transactional information
EPC model (SAP, ARIS, etc)
YAWL model (executable workflow model)
Conversions/exports/imports

- ARIS – ARIS PPM
- BPEL 1.1 (WebSphere/Oracle)
- YAWL
- CPN Tools
- Petrify
- Woflan
- Heuristics nets
- ...
about 30 mining plug-ins!
Social network analysis

**Betweenness Centrality**

<table>
<thead>
<tr>
<th>Node Name</th>
<th>Betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pam</td>
<td>0.25</td>
</tr>
<tr>
<td>Sara</td>
<td>0.25</td>
</tr>
<tr>
<td>John</td>
<td>0.25</td>
</tr>
<tr>
<td>Carol</td>
<td>0.25</td>
</tr>
<tr>
<td>Mary</td>
<td>0.25</td>
</tr>
<tr>
<td>Anne</td>
<td>0.25</td>
</tr>
<tr>
<td>Mike</td>
<td>0.25</td>
</tr>
<tr>
<td>Will</td>
<td>0.25</td>
</tr>
<tr>
<td>Sam</td>
<td>0.25</td>
</tr>
<tr>
<td>Pele</td>
<td>0.25</td>
</tr>
</tbody>
</table>

**Hubs and Authorities**

<table>
<thead>
<tr>
<th>Node Name</th>
<th>HITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pam</td>
<td>0.0025</td>
</tr>
<tr>
<td>Sara</td>
<td>0.0025</td>
</tr>
<tr>
<td>John</td>
<td>0.0025</td>
</tr>
<tr>
<td>Carol</td>
<td>0.0025</td>
</tr>
<tr>
<td>Mary</td>
<td>0.0025</td>
</tr>
<tr>
<td>Anne</td>
<td>0.009999999999997</td>
</tr>
<tr>
<td>Mike</td>
<td>0.009999999999997</td>
</tr>
<tr>
<td>Will</td>
<td>0.0025</td>
</tr>
<tr>
<td>Sam</td>
<td>0.0025</td>
</tr>
<tr>
<td>Pele</td>
<td>0.0025</td>
</tr>
</tbody>
</table>
Decision point analysis builds a decision tree for each choice.
Performance analysis
Discovering patterns
Conformance Checking

“world”
- business processes
- people
- machines
- components
- organizations

models
- analyzes

supports/
- controls

software
- system
- records
- events, e.g.,
- messages,
- transactions,
- etc.

specifies
- configures
- implements
- analyzes

process/
- system
- model

event
- logs

conformance

discovery

verification
Comparing the discovered model with the log (f=1)
Adding deviations to the log \( (f=0.89) \)

Diagram:
- Time-out X (complete)
- Get review X (complete)
- Decide (complete)
- Invite additional reviewer (complete)
- Accept (complete)
- Reject (complete)
LTL checker plug-in
REALITY CHECK AHEAD
TU/e

Goal of ProM: Complete support

Staffware
FLOWer
Websphere
YAWL
ADEPT
ARIS PPM/SIM
Outlook
Caramba
SAP
PeopleSoft
InConcert
AM MQSeries
CPN Tools
CVS
Oracle BPEL
UML SD
Company specific systems
...
Reality Check

• Process mining on structured/administrative workflow-like logs is relatively easy.
• However, let us look at two extreme logs:
  – A log from a hospital with information on treatments, complications, and diagnoses.
  – A log from a manufacturer of professional systems with information on system tests.
First example: Hospital data

- Information on treatment, complication, and diagnosis events.
- Data:
  - 2712 cases (all unique)
  - 29258 events
  - +/- 10.8 events per case
  - 264 different events (activities)
<table>
<thead>
<tr>
<th>Model element</th>
<th>Event type</th>
<th>Occurrences (absolute)</th>
<th>Occurrences (relative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_Perifeer infuus</td>
<td>start</td>
<td>2837</td>
<td>9.696%</td>
</tr>
<tr>
<td>B_Maagsonde</td>
<td>start</td>
<td>2430</td>
<td>8.305%</td>
</tr>
<tr>
<td>B_Beademing</td>
<td>start</td>
<td>2187</td>
<td>7.475%</td>
</tr>
<tr>
<td>B_Catheter a Demeure</td>
<td>start</td>
<td>2096</td>
<td>7.164%</td>
</tr>
<tr>
<td>B_Basiszorg</td>
<td>start</td>
<td>2010</td>
<td>6.87%</td>
</tr>
<tr>
<td>B_Arterie lijn op nk</td>
<td>start</td>
<td>2002</td>
<td>6.843%</td>
</tr>
<tr>
<td>B_O2 masker/slang</td>
<td>start</td>
<td>1954</td>
<td>6.679%</td>
</tr>
<tr>
<td>B_Thoraxdrain</td>
<td>start</td>
<td>1863</td>
<td>6.367%</td>
</tr>
<tr>
<td>Disease</td>
<td>Start</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>C_N Phrenicus Paralyse</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>C_TIA</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>B_Horizontaal</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>C_Cholecystitis, acalc</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>C_Decubitus hak st. 3a</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>C_Druk necrose elders</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>B_Decubitus zorg stadium 3b</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>C_Haemolyse</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>B_Decubitus zorg stadium 4b</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>B_Isolatie Beschermend</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>B_Donor Weefsel</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>C_Polyurie (&gt;40ml/kg/24u)</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>C_Decubitus overig st. 3a</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
<tr>
<td>C_Intra-peritoneaal Abces</td>
<td>1</td>
<td>0,003%</td>
<td></td>
</tr>
</tbody>
</table>
Selection: Care after heart surgery

- Data
  - 874 cases (all unique)
  - 10478 events
  - 181 different events (activities)
Second example: Test data from high-tech system manufacturer

• Information on testing process of complex professional machines
• Data:
  – 24 comparable cases
  – 154966 events
  – +/- 6450 events per case
  – between 2820 and 16250 events per machine
  – 720 different events (start/complete activities)
Helicopter view
Average time spent in job-steps (aggregated events)
Mining just the complete events (# 360)…
Common activities (#70)
Job step level
Conformance checker (reference model – job steps)
Discovered models fit better than reference model

<table>
<thead>
<tr>
<th>Machine ID</th>
<th>Fitness with respect to the reference process model</th>
<th>Fitness with respect to the discovered process model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0431</td>
<td>$f = 0.30895045$</td>
<td>$f = 0.75113416$</td>
</tr>
<tr>
<td>0278</td>
<td>$f = 0.38491702$</td>
<td>$f = 0.82790506$</td>
</tr>
<tr>
<td>0185</td>
<td>$f = 0.37574032$</td>
<td>$f = 0.7171429$</td>
</tr>
<tr>
<td>0466</td>
<td>$f = 0.35643995$</td>
<td>$f = 0.74496365$</td>
</tr>
<tr>
<td>0391</td>
<td>$f = 0.38410193$</td>
<td>$f = 0.72710335$</td>
</tr>
<tr>
<td>1722</td>
<td>$f = 0.33359805$</td>
<td>$f = 0.7599035$</td>
</tr>
<tr>
<td>1694</td>
<td>$f = 0.39662793$</td>
<td>$f = 0.7821204$</td>
</tr>
<tr>
<td>1256</td>
<td>$f = 0.40988705$</td>
<td>$f = 0.7436544$</td>
</tr>
<tr>
<td>1343</td>
<td>$f = 0.3985147$</td>
<td>$f = 0.70074475$</td>
</tr>
<tr>
<td>1981</td>
<td>$f = 0.35723096$</td>
<td>$f = 0.6668054$</td>
</tr>
<tr>
<td>1754</td>
<td>$f = 0.401513$</td>
<td>$f = 0.77557445$</td>
</tr>
<tr>
<td>1662</td>
<td>$f = 0.4138763$</td>
<td>$f = 0.76897943$</td>
</tr>
<tr>
<td>1453</td>
<td>$f = 0.40458512$</td>
<td>$f = 0.5956065$</td>
</tr>
<tr>
<td>1298</td>
<td>$f = 0.37758428$</td>
<td>$f = 0.42357332$</td>
</tr>
<tr>
<td>1876</td>
<td>$f = 0.35556892$</td>
<td>$f = 0.7529762$</td>
</tr>
<tr>
<td>1656</td>
<td>$f = 0.36802232$</td>
<td>$f = 0.65629774$</td>
</tr>
<tr>
<td>1099</td>
<td>$f = 0.424476$</td>
<td>$f = 0.67167395$</td>
</tr>
<tr>
<td>1919</td>
<td>$f = 0.33690846$</td>
<td>$f = 0.72738254$</td>
</tr>
<tr>
<td>1348</td>
<td>$f = 0.41031277$</td>
<td>$f = 0.63753587$</td>
</tr>
<tr>
<td>1596</td>
<td>$f = 0.41015995$</td>
<td>$f = 0.58138484$</td>
</tr>
<tr>
<td>1164</td>
<td>$f = 0.37603533$</td>
<td>$f = 0.67173433$</td>
</tr>
<tr>
<td>1032</td>
<td>$f = 0.32361075$</td>
<td>$f = 0.7062931$</td>
</tr>
<tr>
<td>1794</td>
<td>$f = 0.39387232$</td>
<td>$f = 0.7337234$</td>
</tr>
<tr>
<td>1160</td>
<td>$f = 0.40484217$</td>
<td>$f = 0.7697767$</td>
</tr>
</tbody>
</table>
ProM’s cloud chamber miner
ProM’s Frequency abstraction miner
Conclusion

• Reality is different from models!
• The existence of event data enables a wide variety of process mining techniques: discovery and conformance.
• In the context of services there many event logs around!
• ProM supports this (150 plug-ins)
• Although quite successful for "structured processes", "spaghetti processes" remain a challenge (two examples were given).
• Research should aim to address this challenge.
Relevant WWW sites

- http://www.processmining.org
- http://www.workflowpatterns.com
- http://www.workflowcourse.com
- http://www.win.tue.nl/is/
- http://is.tm.tue.nl/staff/wvdaalst