
Nisha Gupta [nisha1345@iiitd.ac.in]

MTech Research Associate, Software Analytics Research Lab (SARL)
www.software-analytics.in
MTech Thesis Evaluation Committee Members

**Thesis Adviser**
**Prof. Ashish Sureka**
Adjunct Faculty at IIIT-Delhi and currently Researcher at Siemens Corporate Research and Technology
*Faculty In-charge, Software Analytics Research Lab (SARL)*

**External Examiner**
**Dr. Shubhashis Sengupta**
Research Lead and Vice President at Accenture

**Internal Examiner**
**Prof. Chetan Arora**
Faculty Member at IIIT-Delhi
Publications


Work in Progress

Kirtika Anand, Nisha Gupta and Ashish Sureka, Utility-Based Control Flow Discovery from Business Process Event Logs – plan to submit to good data management or business process management conference.
Outline

1. Research Motivation and Aim
2. Related Work and Novel Contributions
3. Research Framework and Solution Approach
4. Experimental Dataset
5. Experiments and Results
6. Limitations and Future Work
7. Conclusion
8. References
Software system that manages and executes operational processes involving people, applications, and/or information sources on the basis of process models [1].

Examples:
- Workflow Management (WFM)
- Business Process Management (BPM)
- Enterprise Resource Planning (ERP)
PAIS log events and activities during the execution of a process.

Figure 1: PAIS
Business Process Logs

- Process consists of cases or incidents.
- Case is a record of events that relate to a single executed process instance.
- Events within a case have attributes such as activity, timestamp, actor, etc.
- Activities in event logs can be modeled as sequential and time series data.
<table>
<thead>
<tr>
<th>Case</th>
<th>Datetime</th>
<th>Activities</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM00000004</td>
<td>2013-01-07 08:17:17</td>
<td>Reassignment</td>
<td>27</td>
</tr>
<tr>
<td>IM00000004</td>
<td>2013-09-25 08:27:40</td>
<td>Operator Update</td>
<td>20</td>
</tr>
<tr>
<td>IM00000004</td>
<td>2013-11-04 12:09:37</td>
<td>Operator Update</td>
<td>20</td>
</tr>
<tr>
<td>IM00000004</td>
<td>2013-11-04 12:09:37</td>
<td>Assignment</td>
<td>2</td>
</tr>
<tr>
<td>IM00000004</td>
<td>2013-11-04 12:09:37</td>
<td>Reassignment</td>
<td>27</td>
</tr>
<tr>
<td>IM00000004</td>
<td>2013-11-04 13:41:30</td>
<td>Reassignment</td>
<td>27</td>
</tr>
</tbody>
</table>
Business Process Logs

- Process consists of cases or incidents.

- Case is a record of events that relate to a single executed process instance.

- Events within a case have attributes such as activity, timestamp, actor, etc.

- Activities in event logs can be modeled as sequential and time series data.
### Research Motivation and Aim

### Business Process Logs

<table>
<thead>
<tr>
<th>IncidentID</th>
<th>DateStamp</th>
<th>IncidentActivity_Number</th>
<th>IncidentActivity_Type</th>
<th>AssignmentGroup</th>
<th>KMNumber</th>
<th>InteractionID</th>
<th>IncidentActivity_Type_Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM0000004</td>
<td>2013-01-07 08:17:17</td>
<td>001A3689763</td>
<td>Reassignment</td>
<td>TEAM0001</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>27</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-09-25 08:27:40</td>
<td>001A5544096</td>
<td>Operator Update</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>20</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849980</td>
<td>Operator Update</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>20</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849979</td>
<td>Assignment</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>2</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849978</td>
<td>Reassignment</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>27</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 13:41:30</td>
<td>001A5852941</td>
<td>Reassignment</td>
<td>TEAM0002</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>27</td>
</tr>
</tbody>
</table>
Business Process Logs

- Process consists of cases or incidents.
- Case is a record of events that relate to a single executed process instance.
- Events within a case have attributes such as activity, timestamp, actor, etc.
- Activities in event logs can be modeled as sequential and time series data.
### Business Process Logs

<table>
<thead>
<tr>
<th>IncidentID</th>
<th>DateStamp</th>
<th>IncidentActivity_Number</th>
<th>IncidentActivity_Type</th>
<th>AssignmentGroup</th>
<th>KMnumber</th>
<th>InteractionID</th>
<th>IncidentActivity_Type_Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM00000004</td>
<td>2013-01-07 08:17:17</td>
<td>001A3689763</td>
<td>Reassignment</td>
<td>TEAM0001</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>27</td>
</tr>
<tr>
<td>IM00000004</td>
<td>2013-09-25 08:27:40</td>
<td>001A5544096</td>
<td>Operator Update</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>20</td>
</tr>
<tr>
<td>IM00000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849980</td>
<td>Operator Update</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>20</td>
</tr>
<tr>
<td>IM00000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849979</td>
<td>Assignment</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>2</td>
</tr>
<tr>
<td>IM00000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849978</td>
<td>Reassignment</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>27</td>
</tr>
<tr>
<td>IM00000004</td>
<td>2013-11-04 13:41:30</td>
<td>001A5852941</td>
<td>Reassignment</td>
<td>TEAM0002</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>27</td>
</tr>
<tr>
<td>IncidentID</td>
<td>DateStamp</td>
<td>IncidentActivity_Number</td>
<td>IncidentActivity_Type</td>
<td>AssignmentGroup</td>
<td>KMnumber</td>
<td>InteractionID</td>
<td>IncidentActivity_Type_Number</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------</td>
<td>-------------------------</td>
<td>-------------------------</td>
<td>-----------------</td>
<td>----------</td>
<td>---------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-01-07 08:17:17</td>
<td>001A3689763</td>
<td>Reassignment</td>
<td>TEAM0001</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>27</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-09-25 08:27:40</td>
<td>001A5544096</td>
<td>Operator Update</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>20</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849980</td>
<td>Operator Update</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>20</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849979</td>
<td>Assignment</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>2</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849978</td>
<td>Reassignment</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>27</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 13:41:30</td>
<td>001A5852941</td>
<td>Reassignment</td>
<td>TEAM0002</td>
<td>KM0000553</td>
<td>SD0000007</td>
<td>27</td>
</tr>
</tbody>
</table>
### Research Motivation and Aim

#### Business Process Logs

<table>
<thead>
<tr>
<th>Case</th>
<th>Datetime stamp</th>
<th>Activities</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM0000004</td>
<td>2013-01-07 08:17:17</td>
<td>001A3689763</td>
<td>Reassignment</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-09-25 08:27:40</td>
<td>001A5544096</td>
<td>Operator Update</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849980</td>
<td>Operator Update</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849979</td>
<td>Assignment</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849978</td>
<td>Reassignment</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 13:41:30</td>
<td>001A5852941</td>
<td>Reassignment</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 13:41:30</td>
<td>001A5852941</td>
<td>Reassignment</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AssignmentGroup</th>
<th>KMnumber</th>
<th>InteractionID</th>
<th>IncidentActivity_Type_Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEAM0001</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>27</td>
</tr>
<tr>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>20</td>
</tr>
<tr>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>20</td>
</tr>
<tr>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>27</td>
</tr>
<tr>
<td>TEAM0002</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>27</td>
</tr>
</tbody>
</table>
Patterns in data that do not confirm to a well defined notion of normal behavior.
PAIS are vulnerable to System Failures, Malfunctions and Frauds which results in anomalies. Example: identify changes in employee behavior that signal a security breach.

Difficult to identify anomalies due to large volume of event logs.

To identify important changes in customer and application behavior by studying business process event logs.
Anomaly

Dig down to the roots to get the root cause

Anomaly effectively addressed
Challenging for an organization to discover the specific actions that lead to anomalies in business process logs.

Prevent the problem from reoccurring in business process flow.

Improve the process performance, work quality and enhance productivity.
Research Aim

- To investigate Window based and Markovian based techniques for detecting anomalies in business process event logs.

- To apply machine learning techniques such as decision tree classifier to extract rules describing cause of anomalous behavior.

- To interactively explore different patterns of data using advanced visualization techniques such as parallel plot.
To assist a process analyst to analyze decision tree and parallel plot results, thus identifying root cause of anomalous incidents.

To demonstrate the effectiveness of our proposed approach using triangulation study.
Outline

1. Research Motivation and Aim
2. Related Work and Novel Contributions
3. Research Framework and Solution Approach
4. Experimental Dataset
5. Experiments and Results
6. Limitations and Future Work
7. Conclusion
8. References
Related Work and Novel Contributions

Related Work

Calderón-Ruiz et al. [2]
- Novel technique to identify potential causes of failures in business process based event logs.
- Event log divided in two, successful and failed cases.
- Patterns compared using control flow and time perspective.

Heravizadeh et al. [3]
- Conceptual methodology of root cause analysis in business processes.
- Based on definitions of softgoals for all process activities.
- Requires much effort from participant.
Related Work II

- Suriadi et al. [4]
  - Enrich and transform process based logs for root cause analysis.
  - Based on classification algorithms.

- Vasilyev et al. [5]
  - Approach to find cause of delays based on the information recorded in event logs.
  - Application of decision tree learner TILDE.
Bezerra et al. [6]

- Approach based on incremental mining for anomaly detection.
- Cannot deal with longer traces and/or logs with various classes of traces.

Bezerra et al. [7]

- Extend own work in [6].
- Anomaly detection model based on the discovery of an “appropriate process model”.
Related Work IV

- Bezerra et al. [8]
  - Extend own work in [6].
  - Three algorithms: threshold, iterative and sampling for anomaly detection.
  - Based on both conformance analysis and process discovery.

- Bezerra et al. [9]
  - Extend own work in [8].
  - Dynamic threshold algorithm for anomaly detection in logs of PAIS.
Novel Research Contributions

- Detection of anomalous traces in business process event-logs using Window-based and Markovian-based techniques.

- Application of parallel coordinate plots for visualizing the characteristics of anomalous and normal traces.

- Root Cause Analysis on business process data using decision tree and parallel coordinate plot.

- Conducted a study on a real-world dataset (Rabobank Group: Activity log for incidents).
3 Research Framework and Solution Approach
PARIKET

Research Framework and Solution Approach

Architecture Diagram

1. BPI 2014 DATA SET
   - Zoom Shape 1
   - MySQL

2. bpi2014 Tables
   - change_detail
   - incident_activity_detail
   - incident_detail
   - interaction_detail

Input:
- ID: ActivitySeq
  - ID1: 4;6;1;8;3
  - ID2: 7;15;0;17
  - ID3: 22;14;6;13
  - ID4: 3;19;5;4

SEQUENCE OF ACTIVITIES ACCORDING TO TIMESTAMP

Window & Markovian Technique

DETECTING ANOMALIES

OUTPUT

Zoom Shape 2

Anomaly Detected

Zoom Shape 3

Zoom Shape 4

Zoom Shape 5

Zoom Shape 6

Visualisation

Classification

Data Preprocessing

Root Cause, Fraud & Failure Analysis
SEQUENCE OF ACTIVITIES ACCORDING TO TIMESTAMP

ID: ActivitySeq
ID1: 4;6;1;8;3;
ID2: 2;19;5;1;
ID3: 22;14;6;13;
ID4: 3;19;5;4;
SEQUENCE OF ACTIVITIES ACCORDING TO TIMESTAMP

ID: ActivitySeq
ID1: 4;6;1;8;3;
ID2: 2;19;5;1;
ID3: 22;14;6;13;
ID4: 3;19;5;4;
SEQUENCE OF ACTIVITIES ACCORDING TO TIMESTAMP

ID: ActivitySeq
ID1: 4;6;1;8;3;
ID2: 2;19;5;1;
ID3: 22;14;6;13;
ID4: 3;19;5;4;

Window & Markovian Technique

DETECTING ANOMALIES

Anomaly Detected
2

ID: ActivitySeq
ID1: 4;6;1;8;3;
ID2: 2;19;5;1;
ID3: 22;14;6;13;
ID4: 3;19;5;4;

SEQUENCE OF ACTIVITIES ACCORDING TO TIMESTAMP

INPUT

Window & Markovian Technique

DETECTING ANOMALIES

OUTPUT

Anomaly Detected

DATA

-2.32, 100

-0.02, 0.32, 1.00
Anomaly Detection

Data Preprocessing

\[ \begin{align*}
&\begin{array}{c}
-3.2, 1.00 \\
\end{array} \\
&\begin{array}{c}
-0.02, 0.32, 1.00 \\
\end{array}
\end{align*} \]
CLASSIFICATION
CLASSIFICATION
VISUALISATION

CLASSIFICATION
ROOT CAUSE, FRAUD & FAILURE ANALYSIS

VISUALISATION
ROOT CAUSE, FRAUD & FAILURE ANALYSIS

VISUALISATION
Research Framework and Solution Approach

Architecture Diagram

1. BPI 2014 Data Set
   - MySQL
   - Tables:
     - change_detail
     - incident_activity_detail
     - incident_detail
     - interaction_detail

2. Sequence of Activities According to Timestamp
   - ID: ActivitySeq
   - ID1: 4;6;1;8;3;
   - ID2: 2;19;5;1;
   - ID3: 22;14;6;13;
   - ID4: 3;19;5;4;

3. Window & Markovian Technique
   - Detecting Anomalies
   - Output

4. Data Preprocessing
   - Input

5. Classification
   - 0000

6. Visualisation
   - Root Cause, Fraud & Failure Analysis

Anomaly Detected
Outline

1. Research Motivation and Aim
2. Related Work and Novel Contributions
3. Research Framework and Solution Approach
4. Experimental Dataset
5. Experiments and Results
6. Limitations and Future Work
7. Conclusion
8. References
Experimental Dataset

- Business Process Intelligence 2014 (BPI 2014) dataset.
- Large real world data from Rabobank Group Information and Communication Technology (ICT).
- Related to the Information Technology Infrastructure Library (ITIL) process implemented in the Bank.

**Figure 2**: ITIL process
## Dataset Details

<table>
<thead>
<tr>
<th></th>
<th>Interaction_detail</th>
<th>Incident_detail</th>
<th>Incident_activity_detail</th>
</tr>
</thead>
<tbody>
<tr>
<td># Records</td>
<td>147,004</td>
<td>46,606</td>
<td>466,737</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Incident_activity_detail</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>46,616</td>
</tr>
<tr>
<td>Events</td>
<td>466,737</td>
</tr>
<tr>
<td>Start Timestamp</td>
<td>07.01.2013</td>
</tr>
<tr>
<td>End Timestamp</td>
<td>02.04.2014</td>
</tr>
<tr>
<td>Incident Activity Type</td>
<td>39</td>
</tr>
<tr>
<td>Assignment Group</td>
<td>242</td>
</tr>
</tbody>
</table>
Figure 3: Pareto chart showing the distribution of activities and their cumulative count.
Outline

1. Research Motivation and Aim
2. Related Work and Novel Contributions
3. Research Framework and Solution Approach
4. Experimental Dataset
5. Experiments and Results
6. Limitations and Future Work
7. Conclusion
8. References
Choose Incident_activity_detail for anomaly detection.

Attribute Incident_Activity_Type represents the type of activities performed.

39 unique activities. For example Assignment, Status Change, Update etc.

Assign integer number starting with 0 to 38 to these activities.

Order the activities according to increasing order of DateTime Stamp

Model the sequence of activities as univariate.
### Sequential Dataset Conversion

<table>
<thead>
<tr>
<th>IncidentID</th>
<th>DateStamp</th>
<th>IncidentActivity_Number</th>
<th>IncidentActivity_Type</th>
<th>AssignmentGroup</th>
<th>KNumber</th>
<th>InteractionID</th>
<th>IncidentActivity_Type_Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM0000004</td>
<td>2013-01-07 08:17:17</td>
<td>001A3689763</td>
<td>Reassignment</td>
<td>TEAM0001</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>27</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-09-25 08:27:40</td>
<td>001A5544096</td>
<td>Operator Update</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>20</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849980</td>
<td>Operator Update</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>20</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849979</td>
<td>Assignment</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>2</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 12:09:37</td>
<td>001A5849978</td>
<td>Reassignment</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>27</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 13:41:30</td>
<td>001A5852941</td>
<td>Reassignment</td>
<td>TEAM0002</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>27</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 13:41:30</td>
<td>001A5852943</td>
<td>Update from customer</td>
<td>TEAM0002</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>34</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 13:41:30</td>
<td>001A5852942</td>
<td>Assignment</td>
<td>TEAM0002</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>2</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 13:51:18</td>
<td>001A5852172</td>
<td>Closed</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>5</td>
</tr>
<tr>
<td>IM0000004</td>
<td>2013-11-04 13:51:18</td>
<td>001A5852173</td>
<td>Caused By CI</td>
<td>TEAM0003</td>
<td>KM0000553</td>
<td>SD00000007</td>
<td>4</td>
</tr>
</tbody>
</table>

### IncidentActivity_Type_List

<table>
<thead>
<tr>
<th>IncidentID</th>
<th>IncidentActivity_Type_List</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM0000004</td>
<td>27;20;20;2;27;27;34;2;5;4;</td>
</tr>
<tr>
<td>IM0000005</td>
<td>27;20;9;27;20;2;20;32;32;2;32;27;2;1;2;2;33;2;20;27;2;20;34;27;2;34;32;4;5;</td>
</tr>
</tbody>
</table>
Aim is to identify anomalous incidents based on obtained sequences.

Problem is of unsupervised anomaly detection.

Formal representation of the problem is:
- Given a set of $n$ sequences, $S = \{S_1, S_2, \ldots, S_n\}$ find all sequences in $S$ that are anomalous with respect to $S$.

Two algorithms used:
- Window Based
- Markovian Based
Window Based Technique

Training Phase

<table>
<thead>
<tr>
<th>S.No.</th>
<th>IncidentID</th>
<th>Sequence of activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IM0000004</td>
<td>27;20;20;2;27;34;2;5;4;</td>
</tr>
<tr>
<td>2</td>
<td>IM0000005</td>
<td>27;20;9;27;20;2;20;32;2;32;2;32;1;2;2;33;2;20;27;2;20;34;27;2;34;32;4;5;</td>
</tr>
<tr>
<td>3</td>
<td>IM0000011</td>
<td>27;2;27;20;34;2;34;27;2;34;2;20;27;34;34;2;27;2;5;4;</td>
</tr>
<tr>
<td>4</td>
<td>IM0000012</td>
<td>34;27;2;34;4;5;</td>
</tr>
<tr>
<td>5</td>
<td>IM0000013</td>
<td>34;27;34;2;5;4;</td>
</tr>
<tr>
<td>6</td>
<td>IM0000014</td>
<td>34;34;34;34;2;27;30;5;</td>
</tr>
<tr>
<td>7</td>
<td>IM0000015</td>
<td>34;34;2;27;2;27;33;33;2;20;27;2;33;34;2;27;5;4;</td>
</tr>
</tbody>
</table>

- Consider window size \( k=2 \)
- Create an empty dictionary:
  - Key is subsequence of length 2.
  - Value is frequency of subsequence in whole training set.
Window Based Technique

Take a training sample:
IM0000004: 27;20;20;2;27;27;34;2;5;4;

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>27 20</td>
<td>1</td>
</tr>
<tr>
<td>20 20</td>
<td>1</td>
</tr>
<tr>
<td>20 2</td>
<td>1</td>
</tr>
<tr>
<td>2 27</td>
<td>1</td>
</tr>
<tr>
<td>27 27</td>
<td>1</td>
</tr>
<tr>
<td>27 34</td>
<td>1</td>
</tr>
<tr>
<td>34 2</td>
<td>1</td>
</tr>
<tr>
<td>2 5</td>
<td>1</td>
</tr>
<tr>
<td>5 4</td>
<td>1</td>
</tr>
</tbody>
</table>
Similarly for all training samples, we get:

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>27 20</td>
<td>4</td>
</tr>
<tr>
<td>27 27</td>
<td>1</td>
</tr>
<tr>
<td>9 27</td>
<td>1</td>
</tr>
<tr>
<td>2 32</td>
<td>1</td>
</tr>
<tr>
<td>2 34</td>
<td>4</td>
</tr>
<tr>
<td>2 33</td>
<td>2</td>
</tr>
<tr>
<td>2 2</td>
<td>1</td>
</tr>
<tr>
<td>2 5</td>
<td>3</td>
</tr>
<tr>
<td>34 4</td>
<td>1</td>
</tr>
<tr>
<td>34 34</td>
<td>5</td>
</tr>
<tr>
<td>20 27</td>
<td>3</td>
</tr>
<tr>
<td>20 20</td>
<td>1</td>
</tr>
<tr>
<td>34 2</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subsequence</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 27</td>
<td>1</td>
</tr>
<tr>
<td>27 33</td>
<td>1</td>
</tr>
<tr>
<td>27 30</td>
<td>1</td>
</tr>
<tr>
<td>33 2</td>
<td>2</td>
</tr>
<tr>
<td>27 34</td>
<td>3</td>
</tr>
<tr>
<td>34 32</td>
<td>1</td>
</tr>
<tr>
<td>5 4</td>
<td>4</td>
</tr>
<tr>
<td>30 5</td>
<td>1</td>
</tr>
<tr>
<td>33 34</td>
<td>1</td>
</tr>
<tr>
<td>33 33</td>
<td>1</td>
</tr>
<tr>
<td>2 27</td>
<td>7</td>
</tr>
<tr>
<td>32 2</td>
<td>1</td>
</tr>
<tr>
<td>32 4</td>
<td>1</td>
</tr>
</tbody>
</table>
**Window Based Technique**

**Testing Phase:** Calculate anomaly score of a sequence.

Consider threshold \((\lambda) = 2\)

Test sequence, \(S:\)

IM0000017: 34;27;2;34;4;5;

**Step 1:** \# of windows = length of \(S\) + \(k-1\)

Therefore, \# of windows = 6-2+1= 5

**Step 2:** For all subsequences of a sequence

if freq (subsequence) < \(\lambda\),

\[\text{anomalyScore} = \text{anomalyScore} + 1\]

**Step 3:** \(\text{anomalyScore} (S) = \text{anomalyScore}/ \# \text{ of windows}\)
Window Based Technique

Test sequence, $S$:
IM0000017: 34;27;2;34;4;5;

freq(34, 27) = 4 > $\lambda$
freq(27, 2) = 8 > $\lambda$
freq(2, 34) = 4 > $\lambda$
freq(34, 4) = 1 < $\lambda$
freq(34, 4) = 1 < $\lambda$

anomalyScore = anomalyScore+1

count(4, 5) = 2 = $\lambda$

anomalyScore (S) = 1/5= 0.2
Window Based Technique

- Extract overlapping windows of fixed length \( k \) from a given test sequence.

- Assign anomaly score to each window based on a threshold value \( \lambda \).

- Combine anomaly score of all the windows to obtain an anomaly score for the test sequence.
Algorithm 1: Window Based Algorithm (ID, S, k, λ, N)

Data: IncidentID (ID = ID₁,...,IDₙ) and Sequence of activities (S = S₁,...,Sₙ) from table.

Result: Top N Anomalous IncidentID.

1 set windowSize = k, threshold = λ;
2 create a empty dictionary D, D′
3 create an array list anomalousIncidents;
4 foreach IncidentID IDᵢ in ID do
   5    Sᵢ = get the sequence corresponding to IDᵢ
   6    set windowCount = Sᵢ.length - windowSize + 1
   7    foreach j = 1 to windowCount do
         8    read the subsequence (Wⱼ) of length = windowSize starting from jᵗʰ position in Sᵢ
         9    if Wⱼ is not present in D then
                10       add (Wⱼ, 1) as (key, value) pair in D
            else
                11       add (Wⱼ, value + 1) in D
Experiments and Results

Anomaly Detection

Window Based Technique

```plaintext
foreach IncidentID ID_i in ID do
  S_i = get the sequence corresponding to ID_i
  set windowCount = S_i.length - windowSize + 1
  set anomalyScore = 0.0

  foreach j = 1 to windowCount do
    read the subsequence (W_j) of length = windowSize starting from j^{th} position in S_i
    get the (key, value) pair from D corresponding to key = W_j
    if value is less than threshold then
      anomalyScore = anomalyScore + 1

  anomalyScore = anomalyScore / windowCount
  add ID_i, anomalyScore into D'

sort D' according to decreasing anomalyScore
add top N IncidentId into anomalousIncidents
return anomalousIncidents
```
Based on property of short memory of sequences.

It states that the conditional probability of occurrence of a symbol $s_i$ is dependent on the occurrence of previous $k$ symbols within a sequence $S_i$.

$$P(s_i | s_{(i-k)} \ldots s_{(i-1)}) = \frac{\text{freq}(s_{(i-k)} \ldots s_i)}{\text{freq}(s_{(i-k)} \ldots s_{(i-1)})}$$

*Where,*

- $\text{freq}(s_{(i-k)} \ldots s_i)$ is the frequency of subsequence $s_{(i-k)} \ldots s_i$ in sequence $S$,
- $\text{freq}(s_{(i-k)} \ldots s_{(i-1)})$ is the frequency of subsequence $s_{(i-k)} \ldots s_{(i-1)}$ in sequence $S$. 

Markovian Based Technique
Algorithm 2: Markovian Based Algorithm (ID, S, k, N)

Data: IncidentID (ID = ID₁...IDₙ) and Sequence of activities (S = S₁...Sₙ) from table.

Result: Top N Anomalous IncidentID.

1. create a empty dictionary Dₖ, Dₖ₊₁, D'
2. create an ArrayList anomalousIncidents
3. foreach IncidentID IDᵢ in ID do
   
   Sᵢ = get the sequence corresponding to IDᵢ
   set noOfSubsequences = Sᵢ.length - k + 1

   foreach j = 1 to noOfSubsequences do
     
     read the subsequence (Wⱼ) of length = k starting from jᵗʰ position in Sᵢ
     read the subsequence (Wⱼ₊₁) of length = k+1 starting from jᵗʰ position in Sᵢ

     if Wⱼ is not present in Dₖ then
       add (Wⱼ, 1) as (key, value) pair in Dₖ
     else
       add (Wⱼ, value + 1) in Dₖ

     if Wⱼ₊₁ is not present in Dₖ₊₁ then
       add (Wⱼ₊₁, 1) as (key, value) pair in Dₖ₊₁
     else
       add (Wⱼ₊₁, value + 1) in Dₖ₊₁
Experiments and Results

Anomaly Detection

Markovian Based Technique

```python
foreach IncidentID ID in ID do
    $S_i$ = get the sequence corresponding to $ID_i$
    set noOfSubsequences = $S_i$.length - $k$ + 1
    set anomalyScore = 0.0, prob = 0
    foreach $j = 1$ to noOfSubsequences - 1 do
        read the subsequence ($W_j$) of length = $k$ starting from $j^{th}$ position in $S_i$
        read the subsequence ($W_{j+1}$) of length= $k+1$ starting from $j^{th}$ position in $S_i$
        get the $(key_j, value_j)$ pair from $D_k$ corresponding to key = $W_j$
        get the $(key_{j+1}, value_{j+1})$ pair from $D_{k+1}$ corresponding to key = $W_{j+1}$
        $r = (value_j) / (value_{j+1})$;
        prob = prob + log $(r)$;
    prob = prob / noOfSubsequences
    TestSequenceProbability = $e^{prob}$
    anomalyScore = 1 / TestSequenceProbability;
    add $ID_i$, anomalyScore into $D'$
sort $D'$ according to decreasing anomalyScore
add top $N$ IncidentID into anomalousIncidents
return anomalousIncidents
```
Aim is to identify root cause of anomalous incidents.

Data mining technique such as decision tree to extract rules describing anomalous behavior.

Input is required in particular format and of high quality.

Pre-processing brings data in required format and helps in improving accuracy and efficiency of mining process.
## Data Pre-processing

### Interaction_detail

### Incident_detail

### Interaction_Incident

- **Anomalous (Yes/No)**

**Datetime format** ‘yyyy-MM-dd HH:mm:SS’ → ‘hours’
## Data Pre-processing

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Attribute Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIType (Aff)</td>
<td>Nominal</td>
<td>13 distinct type of CI’s. Example: software, storage, hardware.</td>
</tr>
<tr>
<td>CISubType (Aff)</td>
<td>Nominal</td>
<td>64 CI Sub types. Example: client based, server based.</td>
</tr>
<tr>
<td>Priority</td>
<td>Nominal</td>
<td>{1,2,3,4,5}</td>
</tr>
<tr>
<td>Category</td>
<td>Nominal</td>
<td>4 i.e. {Incident, Request for information, Complaint, Request for change}</td>
</tr>
<tr>
<td>Open Time</td>
<td>Date Format</td>
<td>‘yyyy-MM-dd HH:mm:SS’ format. Converted into ‘hours’.</td>
</tr>
<tr>
<td>Anomalous</td>
<td>Nominal</td>
<td>{Yes, No}</td>
</tr>
</tbody>
</table>
Use decision tree for classification.

- Easy to understand by user
- Handles variety of input
- Handles missing values in dataset

Weka is a collection of machine learning algorithms for data mining tasks

- Flexible interface which is easy to use
Consider combination of attributes or parameters as root causes of anomalous incidents which occur on the path from the root to leaf showing anomalous as Yes.
Weka supports decision tree algorithms, such as J48, ADTree, REPTree and many others.

<table>
<thead>
<tr>
<th>Algorithm Used</th>
<th>Markovian Based</th>
<th>Window Based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>J48</td>
<td>96.3</td>
<td>28.2</td>
</tr>
<tr>
<td>ADTree</td>
<td>96.8</td>
<td>19.6</td>
</tr>
<tr>
<td>REPTree</td>
<td>92.9</td>
<td>27.6</td>
</tr>
<tr>
<td>SimpleCart</td>
<td>93.8</td>
<td>26.1</td>
</tr>
</tbody>
</table>
There are 240 anomalous incidents whose incident open time is greater than 381657 (hrs).
There are 67 anomalous incidents whose interaction open time is greater than 384871 (hrs) and incident open time is greater than 384822 (hrs).
Attributes Incident Open Time, Interaction Open Time, Incident Priority and Incident Category appear on the path which leads to anomalous incident leaf nodes.

There is a path consisting of only Incident Open Time which classifies 240 incidents as anomalous.
To facilitate user interaction with data.

To identify usual and unusual trends which are present in data.

Tibco Spotfire is an analytics software that is used to detect patterns and correlations present in the data.

Parallel Coordinate Plot for visualization.
Each attribute/column is represented on X-axis.

Along Y-axis lowest value of attribute is set to 0% and highest value is set to 100%.

Each row in data table is mapped as a line.

<table>
<thead>
<tr>
<th>Monitor Model</th>
<th>Screen Size</th>
<th>Type</th>
<th># Year Warranty</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>19</td>
<td>CRT</td>
<td>1</td>
<td>150</td>
</tr>
<tr>
<td>B</td>
<td>17</td>
<td>TFT</td>
<td>1</td>
<td>295</td>
</tr>
<tr>
<td>C</td>
<td>19</td>
<td>TFT</td>
<td>2</td>
<td>200</td>
</tr>
<tr>
<td>D</td>
<td>20</td>
<td>CRT</td>
<td>3</td>
<td>299</td>
</tr>
</tbody>
</table>
### Experiments and Results

#### Visualization

<table>
<thead>
<tr>
<th>Point</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ic_CIType (Aff)</td>
<td>No Pattern.</td>
</tr>
<tr>
<td>a</td>
<td>Majority have Ir_OpenTime above 381658 (hrs)</td>
</tr>
<tr>
<td>Ic_Priority</td>
<td>No Pattern</td>
</tr>
<tr>
<td>c, d</td>
<td>Anomalous incidents falls in Incident, and Request for information category.</td>
</tr>
<tr>
<td>b</td>
<td>Majority have Ic_OpenTime above 381658 (hrs).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Point</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ic_CIType (Aff)</td>
<td>No Pattern.</td>
</tr>
<tr>
<td>e</td>
<td>Majority have Ir_OpenTime above 381512 (hrs).</td>
</tr>
<tr>
<td>Ic_Priority</td>
<td>No Pattern</td>
</tr>
<tr>
<td>g, h</td>
<td>Anomalous incidents falls in Incident, and Request for information category.</td>
</tr>
<tr>
<td>F</td>
<td>Majority have Ic_OpenTime above 381512 (hrs).</td>
</tr>
</tbody>
</table>
Visualization
Attributes Incident Open Time, Interaction Open Time and Incident Category can be the cause of anomalies.

CIType (Aff), CISubType (Aff) and Priority cannot be the cause of anomalies.
Consist of gathering evidences from multiple sources to validate.

We use publically available dataset and we do not have facts to validate the root cause.

Real source of problem is confidential.
Triangulation Study

Decision Tree Classifier

Parallel Coordinate Plot

Results Compared and Interpreted
### Triangulation Study

<table>
<thead>
<tr>
<th></th>
<th>Observation</th>
<th>Parallel Coordinate Plot</th>
<th>Decision Tree Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Incident Open (IO) Time</td>
<td>IO = 381658 hrs.</td>
<td>381657&lt;IO&lt;= 381658</td>
</tr>
<tr>
<td>2</td>
<td>Incident and Interaction open time (IrO)</td>
<td>383139&lt;IO&lt;387725</td>
<td>383498&lt;IO&lt;=383499</td>
</tr>
<tr>
<td></td>
<td></td>
<td>383139&lt;IrO&lt;387706</td>
<td>IrO&gt;383499</td>
</tr>
<tr>
<td>3</td>
<td>Category</td>
<td>Incident, Request for Information</td>
<td>Incident</td>
</tr>
<tr>
<td>4</td>
<td>Affected CI type and CI subtype</td>
<td>No dependency</td>
<td>No dependency</td>
</tr>
<tr>
<td>5</td>
<td>Closure Code</td>
<td>No dependency</td>
<td>No dependency</td>
</tr>
<tr>
<td>5</td>
<td>Root Causes</td>
<td>Open time of Incident, Interaction open time and category</td>
<td>Open time of Incident, Priority, Interaction open time and category</td>
</tr>
</tbody>
</table>
# Outline

1. Research Motivation and Aim
2. Related Work and Novel Contributions
3. Research Framework and Solution Approach
4. Experimental Dataset
5. Experiments and Results
6. Limitations and Future Work
7. Conclusion
8. References
Limitations and Future Work

- Small size of dataset.
- Real source of the problem is confidential and not known to us or publicly available.
- Apply our approach for root cause analysis in other domains and on data from system prone to failures and frauds.
- Analysis are based on two evidences of triangulation study.
- Improve triangulation study by gathering evidences with more classification and visualization techniques.
Outline

1. Research Motivation and Aim
2. Related Work and Novel Contributions
3. Research Framework and Solution Approach
4. Experimental Dataset
5. Experiments and Results
6. Limitations and Future Work
7. Conclusion
8. References
Conclusion

- Novel approach for identification of anomalous traces and executions from business process event-logs.
- New technique for Root Cause Analysis (RCA) of anomalous traces.
- Experimental results reveal agreement in output from Window-based and Markovian technique.
- Data pre-processing and transformation is needed and impacts the outcome of parallel coordinate plot and decision tree classifier.
- Triangulation study to demonstrate that the proposed approach is effective.
Bedankt

多謝

ขอบคุณ

Спасибо

Thank You

धन्यवाद

شکراً

Merci

Obrigado

Gracias!

多谢

Danke

감사합니다
Outline

1 Research Motivation and Aim
2 Related Work and Novel Contributions
3 Research Framework and Solution Approach
4 Experimental Dataset
5 Experiments and Results
6 Limitations and Future Work
7 Conclusion
8 References


Suriadi Suriadi, Chun Ouyang, Wil MP van der Aalst, and Arthur HM ter Hofstede.  
Root cause analysis with enriched process logs.  

Evgeniy Vasilyev, Diogo R Ferreira, and Junichi Iijima.  
Using inductive reasoning to find the cause of process delays.  

Stephen Muggleton and Luc De Raedt.  
Inductive logic programming : Theory and methods.  

Hendrik Blockeel and Luc De Raedt.  
Top-down induction of first-order logical decision trees.  

Jacques Wainer, Kwanghoon Kim, and Clarence A Ellis.  
A workflow mining method through model rewriting.  
References

Fabio Bezerra and Jacques Wainer.

Fabio Bezerra, Jacques Wainer, and Wil MP van der Aalst.

Fabio Bezerra and Jacques Wainer.

Fabio Bezerra and Jacques Wainer.