Nirikshan: Process Mining Software Repositories to Identify Inefficiencies, Imperfections, and Enhance Existing Process Capabilities

Monika Gupta
monikag@iiitd.ac.in

PhD Advisor: Dr. Ashish Sureka
Industry Mentor: Dr. Srinivas Padmanabhuni

Indraprastha Institute of Information Technology
New Delhi, India
Presentation Outline

- Research Motivation
- Research Aim
- Related Work
- Related Methodology and Proposed Contributions
  - Data Source
  - Technique and Contributions
  - Evaluation
- Preliminary Results
“If one cannot measure it, one cannot improve it”
Research Motivation

Process Mining:

• Extract knowledge from event logs recorded by an information system [1].

• Event logs (e.g. transaction logs) with four fields:

<table>
<thead>
<tr>
<th>CaseID</th>
<th>Event</th>
<th>Timestamp</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

• Tools and framework for process mining:
  ➢ ProM ¹ (Open Source)
  ➢ Disco ² (Commercial)

Software Repositories:

• Artifacts generated by the tools during software evolution and archived for future reference.

• Rich data available.

• Uncover interesting and actionable information for process improvement.

For example:
Issue Tracking System (ITS),
Version Control System,
Code Review etc.
Research Aim

Novel applications of process mining on software repositories for:

• Runtime Process Map Discovery
• Performance Analysis
• Conformance verification
• User behavior pattern investigation
• Enhancement of process mining capabilities of existing tools
## Related Work

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Repository</th>
<th>Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rubin et al.</td>
<td>2007</td>
<td>Subversion logs of the ArgoUML project. (OSS)</td>
<td>Linear Temporal Logic (LTL) Checking (Conformance Analysis), Social Network Discovery, Performance Analysis, Petri Net Discovery</td>
</tr>
<tr>
<td>Akman et al.</td>
<td>2009</td>
<td>Software Configuration Management of industry project.</td>
<td>Analyzed and compared the effectiveness of four process discovery algorithms on software process. Analyzed discrepancies between real time and design time process.</td>
</tr>
<tr>
<td>Knab et al.</td>
<td>2010</td>
<td>ITS of EUREKA project SERIOUS (CSS).</td>
<td>Interactive approach to visualize effort estimation and process lifecycle patterns in ITS to detect outliers, flaws and interesting properties.</td>
</tr>
<tr>
<td>Poncin et al.</td>
<td>2011</td>
<td>aMSN and GCC bug repositories, mail archives, SVM</td>
<td>Combined different repositories for analysis using a prototype, FRASR. Role classification and Bug life cycle construction using ProM.</td>
</tr>
<tr>
<td>Sunindyo et al.</td>
<td>2012</td>
<td>Red Hat Linux ITS</td>
<td>Framework for collecting and analyzing data from bug reporting system, conformance checking to improve process quality.</td>
</tr>
</tbody>
</table>
Technical Challenges

• Mining from multiple perspectives

• Gathering data from heterogeneous sources

• Data incompatibility with process mining tools

• Mining hidden tasks

• Missing clear design process and goals
NIRIKSHAN (Sanskrit word which means ‘to investigate’):
Research framework showing proposed research approach for the research contributions followed by evaluation.
Research Methodology

PRACTITIONER’S SURVEY AND FIELD STUDY

- Validate initial hypothesis
- Verify in-practice process and policies
- Identify research problems

DATA SOURCE

- Issue Tracking System like Bugzilla, JIRA, Mantis
- Peer Code Review System like Gerrit, Rietveld
- Version Control System like SVN and Mercurial
- Project Hosting Platform like Sourceforge, Git

Research Questions that usually process analysts have and can be answered by process mining

Event Log Data

APIs, XML-RPC, JSON-RPC, Web Scraping
NIRIKSHAN (Sanskrit word which means ‘to investigate’): Research framework showing proposed research approach for the research contributions followed by evaluation.
Research Methodology

Single Repository

MySQL RDBMS

CaseID, Timestamp, Activity, Actor

Event Log

Data Integration

PROCESS PERSPECTIVE

ORGANIZATIONAL PERSPECTIVE

Runtime Process Model Discovery

Performance Analysis

Conformance Verification

Joint Cases, and Joint Activity

Handover, and Subcontracting
Research Methodology

![Diagram showing process flow for Customer 1, Customer 2, and Customer 3 with stages of Apply, Verify, Check, Approve, Reject, and Send]
Research Methodology

Technique And Contribution

Research Methodology

Technique And Contribution

Research Methodology

Technique And Contribution

Single Repository

Data Integration

MySQL RDBMS

Event Log

CaseID
Timestamp
Activity
Actor

PROCESS PERSPECTIVE

ORGANIZATIONAL PERSPECTIVE

Runtime Process Model Discovery
Performance Analysis
Conformance Verification
Joint Cases, and Joint Activity
Handover, and Subcontracting
“Do we do what was agreed upon?”

Design Process Model v/s Discovered Runtime Process Model
Preliminary Results


- Event log data of ITS analyzed.
- Case study on open source Firefox browser and Core project.
- Discover process map, self-loops, back-forth, issue reopen.
- Algorithm for conformance verification.

Insights:

- 15 activities in event log
- Component and developer reassignment frequent
- 3-4 events in lifecycle, few with >7 events
- 80% of the cases covered with 2% unique traces
- Total unique traces: Core-1164, Firefox-622

Process Map for Firefox with labels as absolute frequency of transition and activity
Preliminary Results

Bottleneck:
• Worksforme, Wontfix state

Reopened:
• Worksforme, Wontfix, Fixed

Conformance:
• 0.86 for Core, 0.91 for Firefox
• Cause of inconsistency:

  Reported → Assigned
Preliminary Results


• Issue resolution process starting from issue reporting in ITS, followed by patch submission to PCR for review and finally committed to VCS.
• Case study on open source Google Chromium project.
• Discover process map, anti-patterns, bottlenecks, organizational metrics.

Preliminary Results

ITS BUG ID

**Issue 88294**: default printing settings are always "two sided"
- **Status**: Fixed
- **Owner**: kmadh...@chromium.o
- **Closed**: Jul 2011
- **Type**: Bug
- **Pri**: 2
- **OS**: Windows
- **Cr**: Internals
- **M**: 14

| Status: Fixed |
| Owner: kmadh...@chromium.o |
| Closed: Jul 2011 |
| Type: Bug |
| Pri: 2 |
| OS: Windows |
| Cr: Internals |
| M: 14 |

The following revision refers to:

http://arc.chromium.org/viewvc/chrome?view=rev&revision=92154

**Description**

PrintPreview: [WIN] Fix the default duplex print setting.

**Commit**: http://arc.chromium.org/viewvc/chrome?view=rev&revision=92154

**Patch Set 1**
- **Total comments**: 2

**Patch Set 2**: Fixed nit

**Revision**: 92154

**Version Control System**

**VCS Revision ID**: 92154

**PCR Issue ID**: 7285039

**Peer Code Review System**

**ITS Issue ID**: 7285039

PrintPreview: [WIN] Fix the default duplex print setting.

**Commit**: http://arc.chromium.org/viewvc/chrome?view=rev&revision=92154

**Bug**: 88294

**Test**: Please refer to bug report.

**Review URL**: http://codereview.chromium.org/7285039

**Author**: kmadh...@chromium.o

**Date**: Tue Jul 12 12:55:14 2011 UTC (2 years, 10 months ago)

**Changed paths**: 1

**Log Message**:

PrintPreview: [WIN] Fix the default duplex print setting.

BUG=88294

TEST=Please refer to bug report.

**Review URL**: http://codereview.chromium.org/7285039

**PCR Issue ID**: 7285039

**Peercodereview.chromium.org**

**Chromiumcodereview.appspot.com**
Preliminary Results

- Resolution process efficient with high chances of issues getting *Fixed*
- Basic and Composite *anti-patterns* like loops, and information flow detected
- *Bottlenecks* are identified such as control transfer between ITS and PCR
- More social performers are more *active*
- Joint activities helps to identify *generalists and specialists*
- Same performer performs *multiple subsequent activities*
NIRIKSHAN (Sanskrit word which means ‘to investigate’):
Research framework showing proposed research approach for the research contributions followed by evaluation.
Research Methodology

EVALUATION

REAL DATA
Open Source Projects
Commercial Projects

Process Mining Approach

VALIDATION:
Survey professionals

Feedback → Improve
Usefulness

- Visualize the process reality for better improvement
- Streamline the process
- Reduce flow time
- Manage changing workloads
- Efficient task allocation
- Better process maintainability, capability, reliability, efficiency and stability [6][7]

References


9. Thomas Zimmermann, Nachiappan Nagappan, Philip J. Guo, Brendan Murphy. ”Characterizing and Predicting Which Bugs Get Reopened”. In Proceedings of the 34th International Conference on Software Engineering (ICSE 2012 SEIP Track), Zurich, Switzerland, June 2012.

THANK YOU!