Tutorial Proposal: Process Mining Software Repositories

Ashish Sureka  
IIIT-Delhi, India  
ashish@iiitd.ac.in

Atul Kumar  
ABB Corporate Research, India  
atul.kumar@in.abb.com

Girish Maskeri Rama  
Infosys Labs, India  
Girish_Rama@infosys.com

ABSTRACT

The tutorial is on the topic of Process Mining Software Repositories which is an emerging field at the intersection of Process Mining and Mining Software Repositories. Process mining is a sub-field of business process intelligence and consists of mining even log data for the purpose of process discovery, conformance checking or verification and process enhancement. Mining Software Repositories consists of analyzing and mining structured and unstructured data stored in various software archives such as version control systems, issue tracking systems, peer code review systems and mail archives to solve problems encountered by practitioners. The tutorial will be of half-day (3-4 hours) duration and the level will be intermediate. The target audiences for the tutorial will be industry practitioners and researchers in the area of data mining, process mining, software analytics and mining software repositories. The pre-requisites for the tutorial is basic background in data mining and software engineering. The tutorial will be divided into 3 parts and cover topics such as: fundamentals of process mining, familiarity with open-source process mining framework ProM, basics of Business Process Modeling Notation (BPMN), overview of mining software repositories and software analytics, understanding of common software repositories and archives, important applications of mining software repositories, basics of process mining software repositories, techniques and applications.

1. BASIC INFORMATION

Title Process Mining Software Repositories

Duration Half-Day (3-4 Hours)

Level Intermediate

Target Audiences: Industrial practitioners and researchers in the area of Data Mining, Process Mining, Software Analytics and Mining Software Repositories.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

The 20th International Conference on Management of Data (COMAD), 17th-19th Dec 2014 at Hyderabad, India.  
Copyright ©2014 Computer Society of India (CSI).

Pre-requisites: Basic background in Data Mining and Software Engineering

Keywords: Analytics, Data Mining, Process Mining, Software Analytics, Mining Software Repositories, Pattern Recognition, Software Engineering

Learning Outcome: Fundamentals of process mining, familiarity with open-source process mining framework ProM, basics of Business Process Modeling Notation (BPMN), overview of mining software repositories and software analytics, understanding of common software repositories and archives, important applications of mining software repositories, basics of process mining software repositories, techniques and applications.

2. TUTORIAL OUTLINE

Table 1, 2 and 3 displays the tutorial outline. The tutorial is divided into 3 parts. Part 1 is on Process Mining (refer to Table 1). Part 2 and 3 are on Mining Software Repositories and Process Mining Software Repositories respectively (refer to Table 2 and 3). Table 1, 2 and 3 shows the sub-topics and the duration for each of the topic. In Part 1 of the tutorial, we will give an introduction to process modeling and cover basics of process modeling, business process modeling notation, process discovery, conformance checking and mining organizational perspective. We will discuss the goals and purpose of a process model. We will introduce the audiences to BPMN which is a graphical representation for specifying business processes in a business process model. Process conformance checking is an important topic and will be covered in the initial 30 minutes. Process conformance checking is about comparing the design time process model with runtime process model (derived from event logs) and measuring the extent of conformance. In addition to introducing the audiences to control flow perspective, we may talk about organizational perspective (optional topic) depending on the available time [5][6].

ProM is an extensible framework that supports a wide variety of process mining techniques in the form of plug-ins. It is platform independent as it is implemented in Java, and can be downloaded free of charge \(^1\). We will conduct a demo and show how to get started with ProM. We will cover ProM user interface and explain the basic features of the user interface by explaining the different objects in this interface. We will then show how to use ProM to answer some of the common questions that managers have about processes in software repositories.

\(^1\)http://www.promtools.org/
Table 1: PART 1: Process Mining

<table>
<thead>
<tr>
<th>Process Mining</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td><strong>Introduction to Process Mining</strong></td>
</tr>
<tr>
<td><strong>Process Modeling</strong></td>
<td><strong>Business Process Modeling Notation (BPMN)</strong></td>
</tr>
<tr>
<td><strong>Conformance Checking</strong></td>
<td><strong>Mining Organizational Perspective</strong></td>
</tr>
<tr>
<td><strong>2</strong></td>
<td><strong>ProM Demo</strong></td>
</tr>
<tr>
<td><strong>Service time of tasks</strong></td>
<td><strong>Time spent between tasks</strong></td>
</tr>
<tr>
<td><strong>3</strong></td>
<td><strong>DISCO Demo</strong></td>
</tr>
<tr>
<td><strong>Automated process discovery</strong></td>
<td><strong>Statistics and charts</strong></td>
</tr>
</tbody>
</table>

Table 2: PART 2: Mining Software Repositories

<table>
<thead>
<tr>
<th>Mining Software Repositories</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td><strong>Software Repositories and Archives</strong></td>
</tr>
<tr>
<td><strong>Version Control Systems</strong></td>
<td><strong>Issue Tracking Systems</strong></td>
</tr>
<tr>
<td><strong>Peer Code Review System</strong></td>
<td><strong>Mail Archives</strong></td>
</tr>
<tr>
<td><strong>Source Code Repositories</strong></td>
<td><strong>Duplicate Bug Report Detection</strong></td>
</tr>
<tr>
<td><strong>2</strong></td>
<td><strong>Overview of Mining Software Repositories</strong></td>
</tr>
<tr>
<td><strong>Fault Localization</strong></td>
<td><strong>Effort and Contribution Estimation</strong></td>
</tr>
<tr>
<td><strong>Automatic Triaging</strong></td>
<td><strong>Defect Proneness</strong></td>
</tr>
</tbody>
</table>

organizations. The tutorial will cover how users can inspect and clean (or pre-process) an event log in ProM. The tutorial will also cover how to mine the control-flow perspective of process models and how to mine information regarding certain aspects of cases. DISCO\(^2\) is a commercial process mining tool. Depending on the available time, we will demo automated process discovery, process map animation and computing detailed statistics from event logs using DISCO.

Mining Software Repositories (MSR) and Software Analytics is an emerging field and has attracted several researcher's attention. We will give an overview of various software repositories such as Version Control Systems (VCS), Issue Tracking System (ITS), Peer Code Review Systems (PCR), Mail Archives and Source Code Repositories. The we will cover some the important applications of MSR such as Duplicate Bug Report Detection, Fault Localization, Effort and Contribution Estimation, Automatic Triaging, Code Clone Detection and detecting Defect Prone areas in the code [1][7]. The last section (Part 3) consists of process mining software repositories (intersection of process mining and mining software repositories). We will cover topics such as: software development processes, technical challenges in process mining software repositories and some of the applications of process mining software repositories [2][3][4].

3. PRESENTER PROFILES

3.1 Ashish Sureka


Ashish Sureka is a Faculty Member at Indraprastha Institute of Information Technology, Delhi (IIIT-D). His current research interests are in the area of Mining Software Repositories, Software Maintenance and Social Media Analytics. He graduated with MS and PhD degree in Computer Science from North Carolina State University (NCSU) in May 2002 and May 2005 respectively. He has worked at IBM Research Labs in USA and was a Senior Research Associate at the R&D Unit of Infosys Technologies Limited before joining IIIT-D in July 2009. He has received research grants from Department of Information Technology (DIT, Govt. of India), Department of Science and Technology (DST, Govt. of India), published several research papers in international conferences and journals, advising PhD students at IIIT Delhi, graduated several research students, organized workshops co-located with conferences, received best paper award and holds seven granted US patents.

3.2 Atul Kumar

Atul Kumar is a Principal Scientist in the Software Research group at ABB Corporate Research, India. Before ABB, he has worked at IBM Research, Microsoft and Accenture Technology Labs. His research interest are in the areas of Distributed Systems, Software Engineering, Internet Technologies and Data Engineering. He has co-organized workshops and special sessions related Software Engineering and Cloud Computing at various conferences ICSE, ISEC, i-Society etc. He is serving as tutorials co-char at ICIS 2014. Atul holds a masters degree and a PhD in Computer
Science from IIT Kanpur. Atul is a senior member of both IEEE and ACM.

3.3 Girish Maskeri Rama
Girish Maskeri Rama is a senior research scientist at Infosys. He has nearly 15 years of experience in applied research and product development. His research focus is on mining software repositories to provide actionable insights for better software maintenance. Previously, he worked extensively in software metrics and measurement, software refactoring, program comprehension, and model driven software development. Girish has served as reviewer for several conferences such as Mining Software Repositories (MSR), and ISEC. He has published several papers in international journals and conferences such as IEEE TSE, Journal of Systems and Software, Wiley Software: Practice and Experience, ICSE, ICSM, APSEC and ISEC. Girish has also filed several patents (4 of which has been granted) in various areas of software engineering. Girish received his Masters in Computer Science from University of York, UK. and currently pursuing his PhD at IISc Bangalore.

4. REFERENCES