Research Article

Application of Process Mining in Education: A Case Study

Fazila Baloch, Sania Bhatti* and Amirita Dewani

Department of Software Engineering, Institute of Information and Communication Technologies, Mehran University of Engineering and Technology, Jamshoro, Sindh, Pakistan

*Corresponding Author: Sania Bhatti, Department of Software Engineering, Institute of Information and Communication Technologies, Mehran University of Engineering and Technology, Jamshoro, Sindh, Pakistan.

Abstract

Maintaining large information systems, especially the systems of an educational institute, is fundamental and most challenging. The processes like admission or examination cannot afford time delays, inefficiencies, and inaccuracies. Process mining has recently gained enormous research attention and is a growing technology in evaluating the processes of information systems via event log data recorded over numerous time stamps. The goal of process mining is to find, evaluate, fix, and improve real-world business processes based on the behaviour of an information system as recorded in an event log. The main objective of this research work is to apply process mining techniques for examining and analyzing educational processes using real-time log data of the MUET admission information system. The ProM toolkit is used for process model discovery and deviations. The event data was collected from Mehran UET's admission department. The log files comprising 2 months of access log data typically collected during the admission process were examined. The process models were discovered using different plugins of ProM software. The results proved that with the help of event log files, an actual process model can be generated and see the bottlenecks and inefficiencies.

Keywords: Process Mining; Education Institution; Admission Process; Conformance Checking; ProM

Introduction

Process Mining is a latest and rapidly growing research idea within the area of Business Process Management. The main objective of process mining is to identify, analyze and improve processes by obtaining information from event logs of an information system [1]. For Using process mining technology two things are required, one is there should be a concept of process running and second is Log files which keep track of every activity in the system. Data mining (DM) and process modelling, and analysis are connected through Process Mining (PM). Process mining (PM) offers a process-oriented perspective to the Data mining (DM) technique as a sub-discipline of DM [2]. Process mining follows three basic steps: Process discovery or data preparation,

Conformance or pattern discovery, and extension or analysis. Process discovery algorithms are able to (semi)automate the translation of collected event data into a process model in the field of process mining. Conformance checking algorithms also allow us to determine whether the process execution, as recorded in the event data, is consistent with a reference model [3]. Process mining approaches can generate models that characterize the processes at hand from event logs by evaluating frequent patterns. The Alpha process mining technique, for example, may automatically extract a Petri net from the event log that concisely models' behavior [4]. The question of whether the modelled behavior matches the observed behavior is answered by conformance checking. Model enhancement is the process of improving, extending, and optimizing a model using data from event logs [5-7].

Usually the information systems (software) have two kinds of logs: the access log files, and the error log files. The access log keeps track of the requests that come into the web server. This data could include what sites visitors are looking at, the status of requests, and how long it took the server to respond. This research focuses on deducing information from actual process executions, which are captured in these so-called event logs. Basically, it is a case study of an admission portal, aims to analyze and develop process model of processes. so that the process of admission can be improved.

The admission system is a portal designed to provide online admission facility in the undergraduate degree programs. User/ Students must register themselves first in the portal by creating an account and then further process of filling the admission form starts. By logging in to the account, system will generate a bank challan to user/student, user must submit it to bank, and upload scanned printout to the portal. Then adding basic data/ information, educational data/information, uploading a profile picture and scanned documents must be uploaded. After adding the complete information and choosing the preference for self-finance seat, students got a verified status then they are provided with admit slip for Admission test. And this the last step of online submission of admission form. And in between this process if any of information or document is not properly added or blurred picture of document is uploaded then student got an email and a kind of message to upload or add that information again.

The server processes all the requests and reserves the data in log files. And hence we can apply process mining methods and techniques to gain insights into process models. Event data and log data is used as the input files for this research study. For developing model of admission process, ProM software has been used with process mining algorithms. To decide to go for which algorithm, one must know its queries and situation of data. Furthermore, applying any process mining method without any preprocessing is usually ineffective.

Rest of this paper is organized as follows: Review of existing literature is discussed in section 2. Section 3 details phases of research methodology. Process Model Discovery is elaborated in section 4. Section 5 highlights conformance checking and bottlenecks. Results and recommendations are given in section 6 and finally, section 7 concludes the research work.

Literature Review

The field of process mining has been investigated by different scholars worldwide and its applicability, in terms of case studies has been realized in different domains. Research study accomplished in [8] has utilized techniques of process mining technology for the examination of life pattern of a library information resource/ facility. Seattle public library's data has been used as log data. The study constructed a diagrammatic model of life cycle by using ProM software with inductive visual mining plugin.

Work in [9] presented the outcomes of process mining of an actual case study regulated/operated in San Carlo di Nancy hospital in Rome (Italy). Event Log data files are collected from hospital's clinical data stored by hospital management system. Event log data is collected from hospital information system. Three data sets were analyzed, outpatient clinic with 299685 records, emergency room with 22043 records, and hospitalization with 10843 records and then this data is analyzed using process mining tool ProM from different views and presented the results. Different algorithms are used in performance analysis phase, like inductive visual minor, transition system minor, simple log filter etc. Work accomplished in [10] has applied process mining technique to enhance accident rescue methods of fatal Gas explosion accidents in China. They have taken 50 Gas explosion accidents during period 2006-2014, as log data. Disco and ProM both have been employed in their study. For Final Process model, inductive minor algorithm in ProM is used. The of work in [11] was to uncover student's self-regulated studying during an e-learning studies. They applied process mining methods to achieve this. 101 university students were given a course in 1 semester on the moodle 2.0 platform, and they analyzed students pass and fail ratio. Log data was taken from this platform's event log files.

Research work in [12] gives a review of current state of educational data mining and provided a framework for analyzing educational data provided by platforms like moodle learning management systems. The framework will help in taking decisions or decision support systems. To analyze the basic patterns, latent class analysis (LCA) and sequential pattern mining approaches were utilized; for process mining, heuristic and fuzzy approaches were used to obtain workflows and statistics; and, finally, social-network analysis was used to uncover collaborations. They used

the ProM tool's Heuristic Miner to discover the student learning process. And for fuzzy miner, Disco tool is utilized. Study in [1] discusses cyberattacks, examination/evaluation of alarms and the restrictions of attack reveal/discover systems. They have put forward a model using process mining methods to uplift the present attack detection systems, with the help of that model doubtful activity can be detected in real situations. If there will be any intrusion this model will detect that and safe the data. They have used one hospital data for audit and ProM tool was used for process mining. In this survey paper [2] authors discuss about educational process mining and how this technology is applied in education field. They discuss that initial point of process mining is an event log file, which can be any file that contains sequence of activities. Log files can be collected from learning environments like studying management systems, massive open online courses (MOOCs), adaptive hypermedia network/structure etc. There are many PM tools for processing these log files like Disco, ProM, Celonis discovery, Perceptive process mining, XMAnalyzer etc. The authors have discussed the techniques, algorithms of each step of process mining.

Work carried out in [13] gives attention on student's interaction and actions in distinct internet based quiz- based tasks within learning management systems (LMS). The log data of student's behavior was taken from logging system in Moodle 3.1. for data examination disco tool was used and fuzzy minor algorithm for getting process maps. Researchers in [14] have focused on educational process examination and improvement using process mining techniques. They have taken data from a consulting company connected with training of professionals. Their main aim was to examine training processes and improve training process models. The company has 6000 employees who are allowed to choose training courses. The data gathered for analysis has 16260 courses followed by 3440 employees. In this work [5] an e-trade system is analyzed by using process mining techniques. They have targeted an online ticket booking system for analysis and used ProM software for applying process mining algorithms. Log files of online ticket booking system were used as input and fuzzy model and heuristic model of that system was generated using system's event log data.

Though process mining applications has been practiced in different areas [14,15], however the field has gained little attention in educational domain leaving behind a research gap. To the best of knowledge and literature review this is the first study that typically analyzes admission process in an Engineering University using real time event log data.

Methodology

In this work we look at a case study that aims to do analysis and develop process model of MUET's admission portal. The conformance checking process has also been accomplished. To achieve these goals various data analysis and process mining techniques were used.

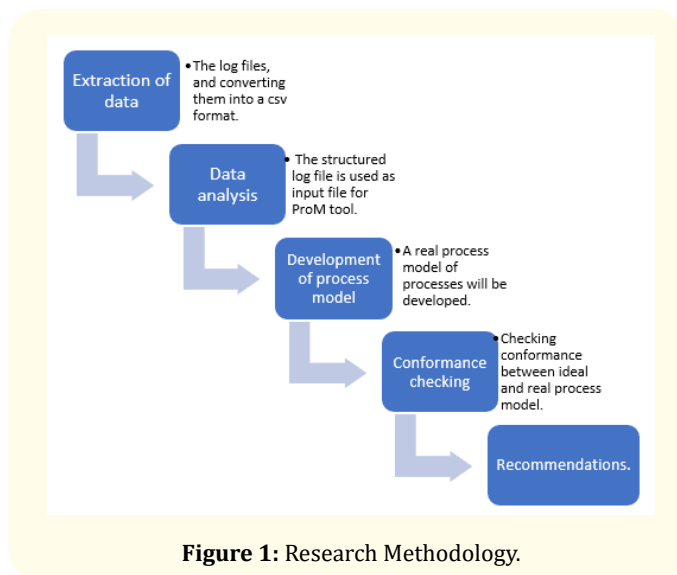The overall layout and phases of research methodology are given below.

Figure 1: Research Methodology.

Two months data were gathered from admission department. As we have gathered two months data, each containing 4453466 and 650 records. Table 1 shows the primary input data. Each row in table 1. represents an event in a case that captures the execution of a specified activity (i.e., an instance of the process). Columns are the attributes associated with the events. Mandatory attributes are the IP which will be our case identifier, the Date and Time column, when the event/activity/request occurred, and the activity or event that was carried out. Other typical attributes are the HTTP response code and the size of the object. Two other columns are not mentioned here, they were the http referrer that refers to the address from the request was generated and the user agent which saves the browser information.

IP	Date and time	Request type and resource being requested	HTTP response code	Size of the object returned to user
. . .	. . .	.	. . .	. . .
1234	[31/ Aug/2021:03:07:35 +0200]	GET/dashboard.php HTTP/1.1	301	250
1235	[31/ Aug/2021:06:47:11 +0200]	GET/forgetPaswordConformation.php HTTP/1.1	301	269
1236	[31/ Aug/2021:07:42:54 +0200]	GET/login.php HTTP/1.1	301	249
1237	[31/ Aug/2021:08:56:50 +0200]	GET/status_candidate_form. php?verifyID=22406 HTTP/1.1	301	280
1238	[31/ Jul/2021:21:26:12 +0200]	GET/data-form.html?error-msg2=Intermediate%20Percentage%20 shouldn%27t%20be%20less%20than%2060. HTTP/1.1	301	330

Table 1: Access log file showing different attributes.

The log files need to be parsed out and from text mining we extracted the information showing in table 2.

Register	Registration on portal
Login	Login
Dashboard	Dashboard of portal
Student form	Student admission form
Basic info student details	Filling of student data/filling basic details
Education details	Adding education details on portal
Upload challan	Upload admission fee challan
Upload docs	Uploading basic documents
Candidate status verified	Checking student form and giving green signal
Download admit slip	Downloading admit slip for entry test

Table 2: Depiction of the functionality of website/portal called as activity or events.

The log files were analyzed and parsed and then converted to csv format. That is because our tool ProM cannot accept the '.gz' format files or unstructured files.

Process model discovery

Aside from event logs, a process model is another common input for process mining approaches. A process model is a blueprint for how a procedure should be carried out. A process model encodes (or should express) the set of permissible executions in a more formal fashion than a textual description in natural language. Using a formal process modelling notation provides multiple advantages, including the ability to evaluate various model quality features, such as the lack of deadlocks. It also enables software tools to reason about the modelled process behavior automatically [3].

We get a process model like the one shown in figure 2 when we apply automated process discovery techniques directly on raw logged click-stream data. Clearly such a complicated process model makes it difficult to achieve the overall purpose of process mining, which is to gain a better understanding of the process.

The model is clearly not human interpretable, that is why, the log file is filtered out and only most frequent events are added. And then the file is analyzed. Figure 3 is showing the most frequent events by ProM tool.
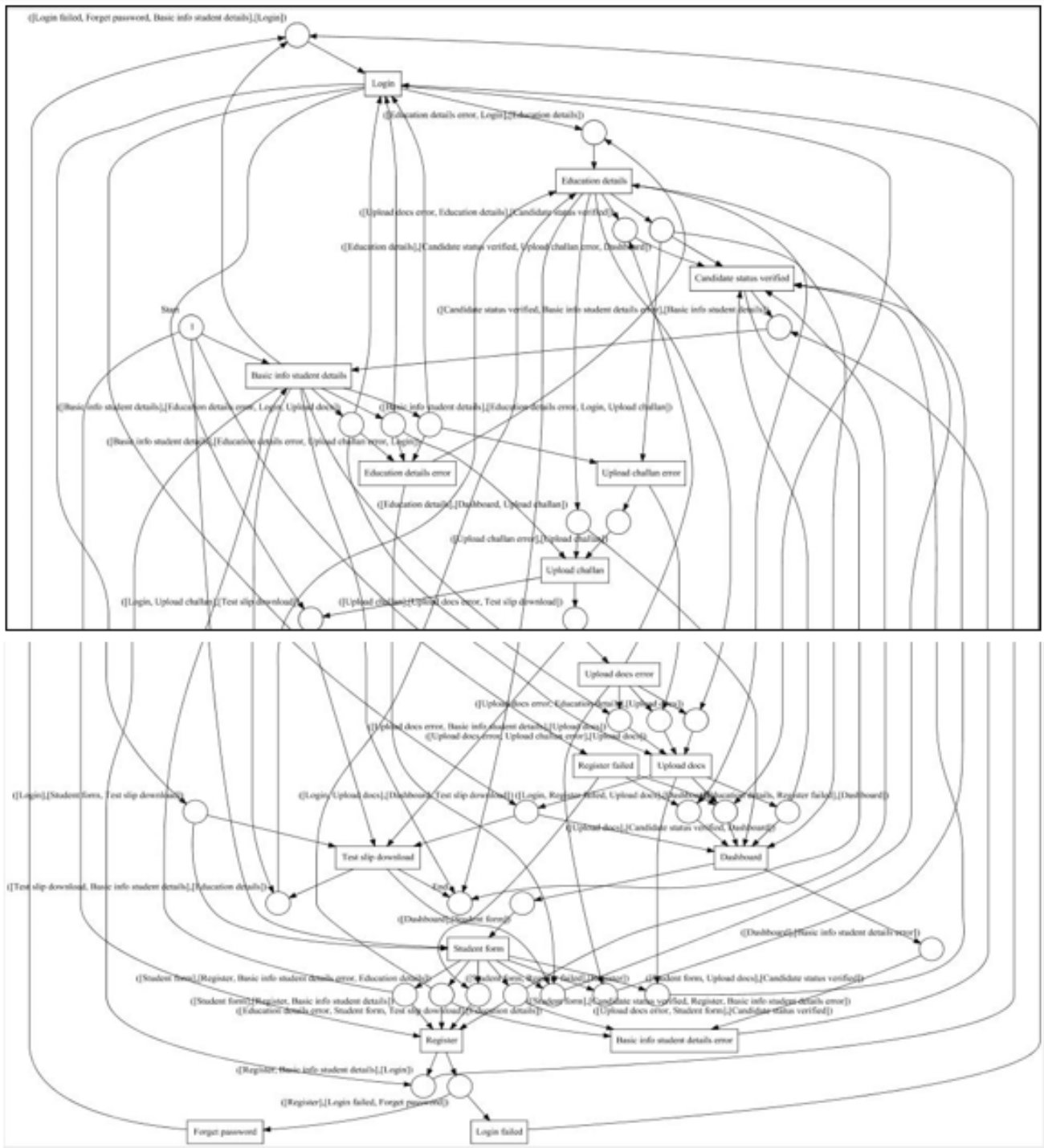
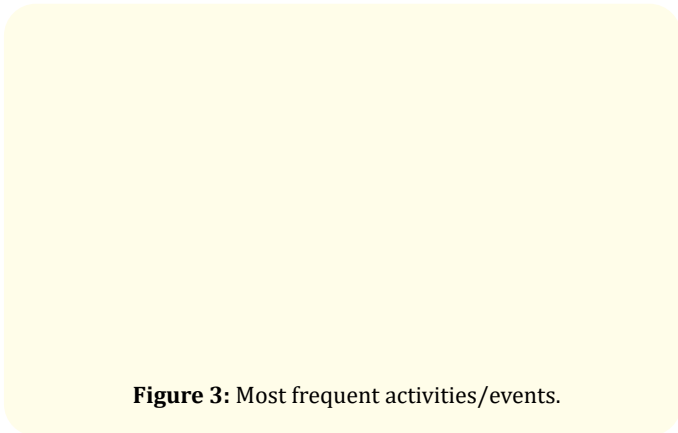**Figure 2:** Process model discovered using process discovery algorithms.

Figure 3: Most frequent activities/events.

Activity graph

After the preprocessing, we have been able to develop a casual activity graph of processes. Figure 4 shows the activity graph, which is how the events are working in a sequence.

Heuristic model

A heuristic model can be used to depict the log's frequency characteristics. The heuristic model is a directed graph with the events as its vertices. The number of traces that contain this event is displayed for each vertex (event). If the matching two occurrences in the event log follow each other directly, two graph vertices are connected by an arc. A frequency parameter (the number of traces having the associated dependency) is assigned to each arc.
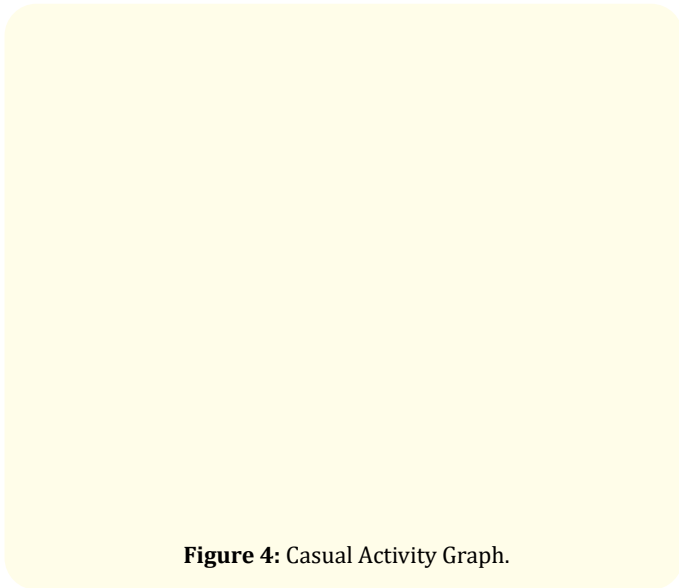

Figure 4: Casual Activity Graph.

The model is discovered/developed using default heuristic parameters initially. Admission process starts with register event, and the student form activity is followed by login event. It is analyzed that mostly occurred events in the log are <Register>, <Login>, <Student Form>, <Basic Info student details>, <education details>, <Upload docs>, and <Test Slip Download>. The discovered Heuristic process net is shown below in figure 5.
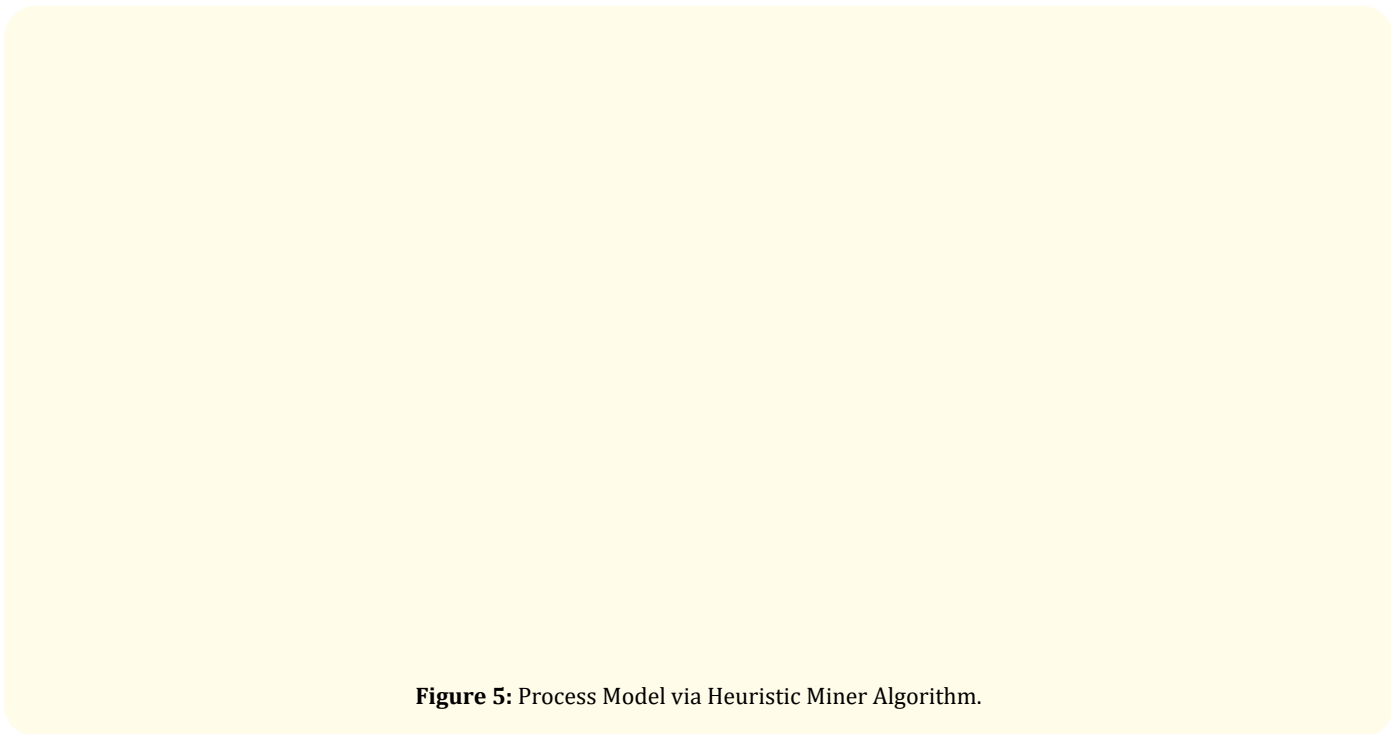

Figure 5: Process Model via Heuristic Miner Algorithm.
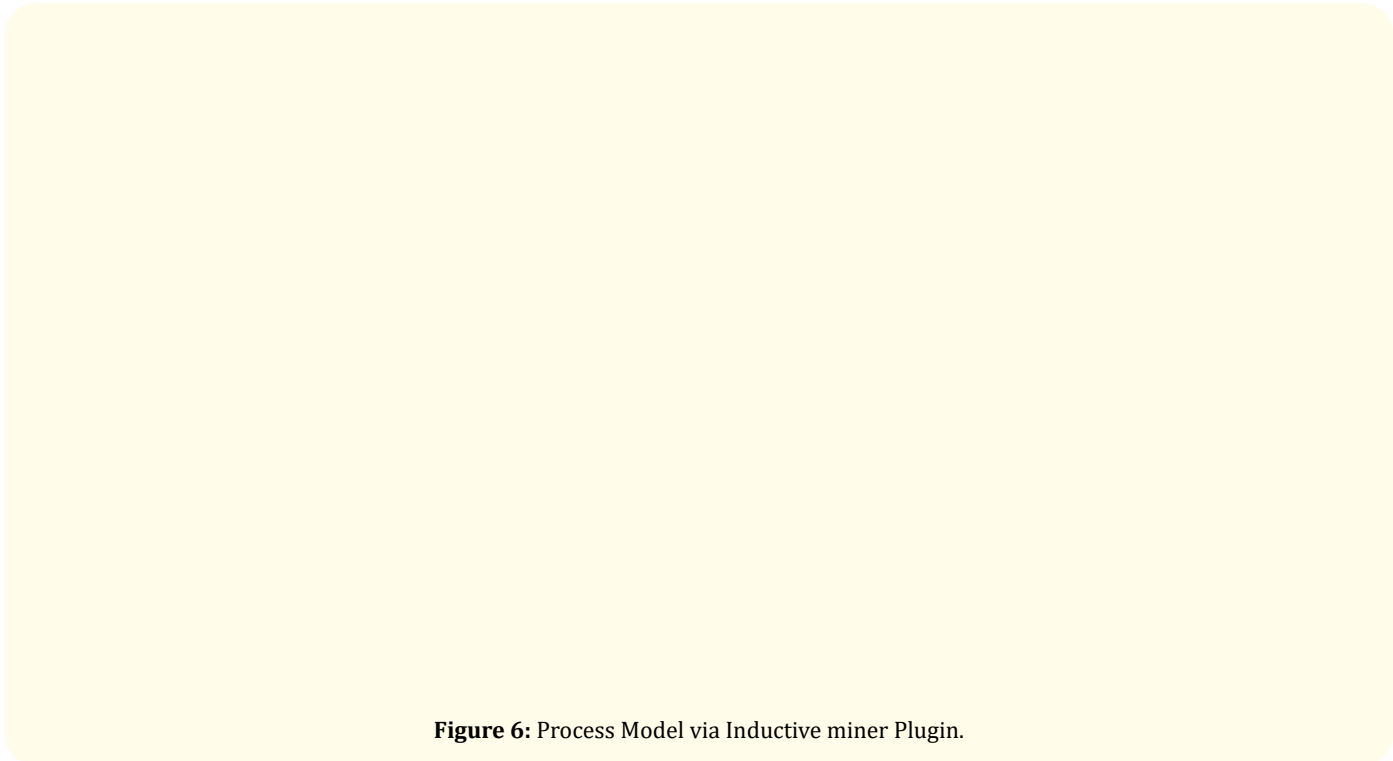
Process discovery via inductive miner

Figure 6: Process Model via Inductive miner Plugin.

Conformance checking

For conformance checking/analysis there are two parameters: one is fitness and other is appropriateness. Appropriateness, i.e. the degree of accuracy with which the process model describes the observed behavior, combined with the degree of clarity with which it is represented, and, Fitness, i.e. the extent to which the log traces can be associated with valid execution paths specified by the process model.

As the portal does not have a documented or formal process model, there was just an idea that how this portal should work, so by doing detailed analysis and from developed process models we have seen the differences/bottlenecks in real processes and some processes who took more user's time.
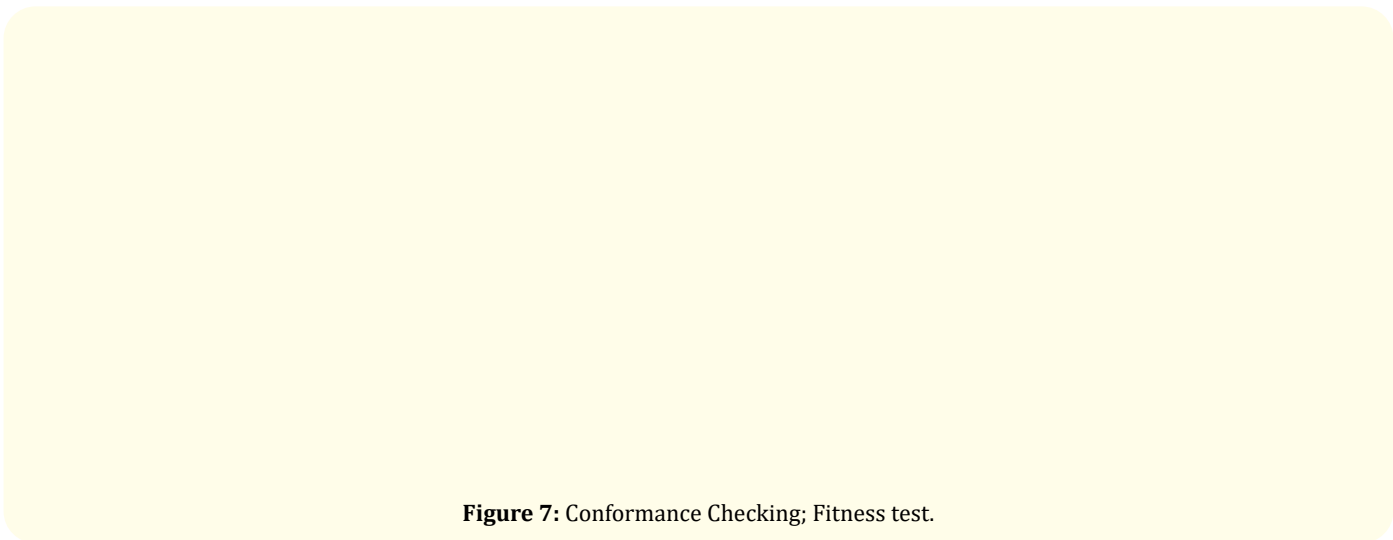

Figure 7: Conformance Checking; Fitness test.

For checking fitness, the log is replayed in the model as shown in Figure 7. The figure 7 shows higher frequency activities got darker blue color, so the activities <register>, <login>, <dashboard>, <student form>, <Basic info student details>, <education details>, <Upload challan>, <upload docs>, <candidate status verified>, and <test slip download> are executed more often. The green/purple bar at the bottom of each transition indicates the percentage of correct/incorrect executions. While an area is highlighted yellow, it means there was an occurrence in the log that couldn't be explained when a token was present. The red boarder means model move is greater than zero and green color border shows there is no model move occur, and the frequency of synchronous move is shown.

The element statistics shows that out of 543, model move value is 289. And the global statistics results show trace fitness is 60%. As during data preprocessing phase, we filtered the log and choose only 545 records, which is less than 40% of most common events remained.

Results and Recommendations

The analysis of process models shows that users/students often forget their passwords as forget password activity seen most in both models also many students got errors in submitting the documents to the portal as Upload docs error event is also seen in log file. Most of the time students leave the portal after registration process and fill the student form later that day or next day. As we have gathered two months data, each containing 4531 and 650 records, after filtering the log data, 544 records were chosen in data preprocessing, because only these selected events were seen repeatedly in log file of admission portal. So, the process model would be same if many other cases were included. Due to heavy traffic on portal events/activities seems taken more time than usual.

As a result of this research following recommendations are given:

Due to heavy traffic students are unable to upload the required documents in one click.

- This step can also be removed, the reason is that user's most time wasted due to heavy traffic and second thing document verification is done manually so no point of uploading all the documents.

- Improve the bandwidth of the channel between user interface and portal server or database.

- There is no need to upload the challan document, administration can get the status of submitted challans by banks. Student just need to enter the challan number.

- Hardware performance must also be improved.

- Administration must update the content of the website/ portal.

- Forget password activity is seen most in the log data, students must get a msg on their phone about their password.

- The user cannot delete the account if he/she has registered himself with wrong CNIC, so delete account option should be there.

Conclusion

The extensive use of process mining in industry has revealed plenty of new difficulties. One of which is applying this technique in education department. The analysis methods were general, and the problems are typical for a portal system. Process mining can help in improving the business processes because the analysis of the system has been done via real event log data. Since the log data captures the user activities, hence from this kind of data, an actual behavior of the system can be analyzed and improved. The crucial question was how to make this e-portal a better version so that student should not feel it difficult or annoyed via filling the admission form online. Event Log files were main data for this work. For analyzing a system, data preprocessing step is most important as the system captures every click on the system by user, therefore there was repetition of events due to which more than 60% of data was eliminated in log filtering. The process mining techniques are more appropriate for analyzing and improving the processes. It is critical to obtain the help of a domain specialist while analyzing procedures. The results must be interpreted to assess the situation. The findings obtained by software are worthless without adaptation and judicious method selection for a specific circumstance. Interpretation is a crucial phase as well. To interpret the findings correctly, one must exercise extreme caution. The correctness of the evaluation results, for example, is highly dependent on the software, data, and settings employed. The results obtained in this study will be highly beneficial if applied in real time setting.

Bibliography

1. V P Mishra., *et al.* "Analysis of alarms to prevent the organizations network in real-time using process mining approach". *Cluster Computing* 22 (2019).

2. A Bogarín., *et al.* "A survey on educational process mining". *WIREs Data Mining and Knowledge Discovery* 8.1 (2018): e1230.

3. S J van Zelst., *et al.* "Event abstraction in process mining: literature review and taxonomy". *Granular Computing* 6.3 (2020): 719-736.

4. P He., *et al.* "An Evaluation Study on Log Parsing and Its Use in Log Mining". presented at the 2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN) (2016).

5. A Mitsyuk., *et al.* "Using Process Mining for the Analysis of an E-trade System: A Case Study". *Business Informatics* 29 (2014): 15-27.

6. W Aalst., *et al.* "Auditing 2.0: Using Process Mining to Support Tomorrow's Auditor". *Computer* 43 (2010): 90-93.

7. A Bogarín., *et al.* "A survey on educational process mining". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.1 (2018): e1230.

8. M Dorrer., *et al.* "Analysis of the life cycle of a library information resource using Process Mining technology". in *Journal of Physics: Conference Series* 1582.1: IOP Publishing (2020): 012024.

9. S Agostinelli., *et al.* "Supporting Governance in Healthcare Through Process Mining: A Case Study". *IEEE Access* 8 (2020): 186012-186025.

10. Z He., *et al.* "A process mining approach to improve emergency rescue processes of fatal gas explosion accidents in Chinese coal mines". *Safety Science* 111 (2019): 154-166.

11. R Cerezo., *et al.* "Process mining for self-regulated learning assessment in e-learning". *Journal of Computing in Higher Education* 32.1 (2020): 74-88.

12. G Ozdagoglu., *et al.* "An application framework for mining online learning processes through event-logs". *Business Process Management Journal* 25 (2018).

13. L Juhaňák., *et al.* "Using process mining to analyze students' quiz-taking behavior patterns in a learning management system". *Computers in Human Behavior* 92 (2019): 496-506.

14. H Awatef., *et al.* "Analyzing and Improving Educational Process Models using Process Mining Techniques" (2015).

15. M Arias., *et al.* "Mapping the patient's journey in healthcare through process mining". *International Journal of Environmental Research and Public Health* 17.18 (2020): 6586.