Examining Diagnosis Paths: A Process Mining Approach

Tenzin Doleck, Amanda Jarrell, Eric G. Poitras, Maher Chaouachi, Susanne P. Lajoie

1McGill University, Montreal, Canada
{tenzin.doleck, amanda.jarrell, maher.chaouachi}@mail.mcgill.ca; susanne.lajoie@mcgill.ca
2University of Utah, Salt Lake City, Utah
eric.poitras@utah.edu

Abstract—This paper is motivated by two observations on computer-supported education: First, there has been growing availability, rapid proliferation, and increased diversity of learner-system educational data. Second, advances in learning analytics and data mining have facilitated and spawned a variety of novel investigations using such data. Driven by these complementary trends, the present work is geared towards exploring knowledge-based discovery approaches in understanding learner-system usage data. More specifically, with an eye toward tracing and comprehending learner behaviors in a medical intelligent tutoring system, we explore the utility of Process Mining, in understanding the problem solving trajectories of students in a medical computer-based learning environment.

Keywords—process mining, intelligent tutoring systems, computer-based learning environments, data mining

I. INTRODUCTION

Much interest has centered on understanding how learners solve ill-structured problems such as diagnosing clinical cases. There has been considerable interest in this topic, however, less is known about the way that novices diagnose clinical cases. Previous studies on clinical reasoning have largely focused on diagnosis correctness. However, the particular problem solving pathways learners take to diagnose a patient is poorly understood. Since the process of clinical reasoning is linked to clinical uncertainty and correctness, there is a need to observe, examine, and better understand the way that learners arrive at a solution. Examinations of learner actions and behaviors will be beneficial to our understanding of clinical reasoning, and more broadly of problem solving. This is an important directive as the way that learners solve a problem is of immense importance to the learning process. A deeper look into such fine-grained information, i.e., learner actions involved in diagnosing a case, may present a portrait of solutions that reveal meaningful insights about learner behaviors.

Knowledge-based discovery methods can be used to accurately capture processes used in real-time, across cases that range in complexity. However, there are few studies that have investigated problem-solving process in computer-based learning environments (CBLEs) or simulation environments. Exploring medical problem solving in a CBLE has several distinct advantages over traditional paper and pencil assessment situations. CBLEs can concurrently track the decisions made to reach a final diagnosis. Thus, CBLEs allow for a pure assessment of clinical reasoning that is independent of retrospective self-report or memory biases. The few studies that have investigated the diagnostic process using CBLEs have been primarily concerned with authentic methods for clinical assessment [1-2]. In these studies it was found that case-specificity (i.e. notion that each clinical cases is unique in terms of, for example, presentation and medical knowledge) obstructed the consistency of student performance on the CBLE clinical problems [1] and in simulated clinical reasoning environments [2]. These findings call for a different approach to evaluating how medical trainees solve clinical cases when cases differ in complexity and in domain knowledge. One alternative to simply evaluating final performance metrics of medical students is to map their problem solving process in order to determine the clinical reasoning strategy they used and how this can predict their performance.

CBLEs provide affordances that make it possible to capture and track learner behaviors [3]; one of the most common ways to track learners’ actions in CBLEs is via log-files. Data mining methods can then be used to analyze the learner-system usage data. In this study, we consider clinical reasoning in a computer-based learning environment from a process perspective; thus, drawing on this perspective we argue that the learner actions involved in solving a clinical case can essentially be viewed in terms of steps or actions in a process. Thus, we employ process mining [4] to infer diagnosis paths.

The paper is organized as follows: (1) the first section presents a brief overview of the computer-based learning environment used in the study, namely, BioWorld; (2) the second section outlines the methods; (3) the third section outlines the procedure for process mining; (4) the fourth section highlights the findings of using process mining; and (5) the final section summarizes the present study.

II. BIOWORLD

BioWorld (Fig. 1) is a CBLE for the medical domain. In BioWorld, novice physicians learn clinical reasoning by diagnosing virtual patient cases through the identification of relevant evidence/symptoms, ordering lab-tests, seeking help via the embedded library, and reasoning about the nature of the underlying disease [5-6]. After submitting their final diagnosis, learners receive feedback based on an aggregated expert solution. The system was created using a cognitive apprenticeship framework [7] where learners practice realistic
clinical reasoning tasks and are scaffolded in the context of their learning with expert models.

![BioWorld Interface](Image)

**III. METHOD**

The following subsection highlights the participant profile, the experimental procedure, and the details of the log files generated by the BioWorld system.

A. **Participants**

The participants for this study were recruited through the university advertisements and faculty of medicine newsletter. A total of 30 undergraduate students agreed and participated in the study. The participants were compensated $20 at the completion of the study. The sample comprised of 19 women (63%) and 11 men (37%), with an average age of 23 (SD = 2.60). All 30 participants were registered in the same classes at a large Northeastern Canadian Research University.

B. **Procedure**

Participants were tasked with solving three virtual patient cases in BioWorld on an individual basis for a total duration of 2 hours. While solving the patient cases, participants also engaged in thinking aloud (i.e. externalizing their mental reasoning). The three virtual patient cases were of different levels of difficulty: Amy (easy), Cynthia (hard), and Susan-Taylor (medium). The correct diagnosis for the cases Amy, Cynthia, and Susan-Taylor was diabetes mellitus (type1), pheochromocytoma, and hyperthyroidism respectively.

C. **Measures: log file**

Each user action is logged by the BioWorld system. Information saved in the log file includes the attempt identifier (participant and case ID), a timestamp of each action, the BioWorld space (e.g., chart), the specific action taken (e.g., add test), and details of the action (e.g., Thyroid Stimulating Hormone (TSH) Result: 0.2 mU/L). The BioWorld system also logs the case summaries written by the user [13]. There are three types of performance metrics that can be extracted from the log files, including, diagnostic efficacy (e.g., accuracy and count of matches with experts), efficiency (e.g., number of tests ordered and time to solve the case), and affect (e.g., confidence).

IV. **PROCESS MINING**

Advances in data mining have produced new and powerful means for examining learning data. Nevertheless, there are some data mining techniques that have not yet been widely implemented in educational research. Process mining, a data mining technique that uses “event data to extract process-related information” [4: 1] has been frequently and increasingly used for investigating and understanding process data in business contexts and has been a staple of business process research. Since there is a paucity of research mining clinical reasoning paths of novice learners, the work we present here leverages the success of process mining observed in business process research and represents an emergent contribution. It is our contention that process mining is a promising method for mining trace data to examine the usage data from CBLEs to explain latent mechanisms that mediate diagnostic performance, and the case-specificity effect found in diagnostic reasoning. According to Rozinat [8], the “core functionality of process mining is the automated discovery of process maps by interpreting the sequences of activities” [8: 3]. We employ the Disco Miner Software [10], based on the framework of Fuzzy Miner, to generate process maps from the BioWorld log files.

Rozinat [9] specifies that the minimum requirements for an event log are Case ID, Activity, and Timestamp; the BioWorld log data meets these requirements. To mine the data, the necessary pre-processing steps were conducted: check for missing values, separation of data for the three cases, and correct mapping of Case ID, Activity, and Timestamp data. The BioWorld log file was imported into the Disco tool and a process model (as a process map) was generated for the data. The Disco tool generates a frequency-based process map, which enables inspection of the process flow between actions; the “process flows …are automatically reconstructed (“discovered”) based on the sequence and timing of the activities” [9: 52] in the log data. Thus, the generated process map illustrates both the order in which the actions have been performed and the relations between the actions (in terms of directionality). In the process map, the triangle symbol and the stop symbol represent the start and end of the process respectively. Each specific action is housed in a box and an arrow marks the process flow between actions. The numbers juxtaposed on the arrows (thicker arrows associated with higher frequencies) and in the boxes (different colors associated with different frequencies) are the absolute frequencies for the transitions and the instances of the actions respectively. The Disco tool allows control of the level of detail presented in the generated maps via two slider controls: Activities and Path sliders. The slider values can be set between 0% and 100%; setting the sliders at 0% shows only the most frequent actions and at 100%, all the actions are revealed. Setting the two sliders to a low value results in the most frequent activities and paths. To allow both interpretability and focus, we decided to set low values for both the slider values, thus giving the most frequent activities and paths for meaningful interpretation.
Fig. 2. Process Model for the Amy Case – Frequency of Action Transitions
Fig. 3. Process Model for the Amy Case- Performance in terms of time
V. FINDINGS

The goal of this preliminary analysis was to present process mining as an approach to understanding learner behaviors with regard to learner problem solving processes; thus, the intention of the findings is to present a global discussion on the utility of process mining. This exercise in general has importance for understanding learner behaviors in clinical reasoning. For the purpose of illustration, we present the generated process maps for one particular case—the Amy case (Figs. 2 & 3).

The approach employed in this case (Fig 2) begins with adding evidence, followed by two distinct blocks of actions. The first block revolves around the formulation of hypotheses (hypotheses formulation block), which begins with adding hypothesis and ends with the final prioritization and submission of summary. After adding hypothesis, the most likely actions include: linking evidence and changing hypothesis conviction. The second block includes two sets of behaviors: ordering lab tests and seeking help (test and help block). Here the relationship between lab tests and help-seeking assumes a reciprocal relationship: ordering of lab tests is followed by help-seeking, and after seeking help, learners tend to order lab tests. For the tail end of the process, a similar pattern of actions was seen across all three cases (prioritization, categorization, and so on); this is because all learners must move through this sequential series of actions in the BioWorld interface.

Along with providing a way to visualize the instances of actions and links between actions, process mining also affords a way to consider the time factor in transitions between actions. In Fig. 3, instead of the instances, the time taken in each action and the time taken between the action transitions are displayed. The examination of the process model incorporating the time factor helps reveal the time spent on each specific activity and the time taken to move from one step to another while arriving at a final solution.

VI. CONCLUSION

Many rich possibilities for examining learner behaviors exist; the process mining approach employed in this study represents a novel way of investigating learner behaviors. The findings substantiate process mining as a useful approach for tracing and uncovering learner behaviors in a medical CBLE. The process mining approach employed in this paper represents our ongoing efforts to examine and understand learner behaviors so as to provide a more coherent picture about clinical reasoning [11-12]. Furthermore, our research adds to the growing area of literature on leveraging data mining techniques for improving insights from educational data.

REFERENCES


