Mining Sequential Patterns of Students' Access on Learning Management System

Leonard K.M. Poon(✉), Siu-Cheung Kong, Michael Y.W. Wong, and Thomas S.H. Yau

Department of Mathematics and Information Technology,
The Education University of Hong Kong, Hong Kong SAR, China
{kmpoon,sckong,mywwong,shyau}@eduhk.hk

Abstract. Novel pedagogical approaches supported by digital technologies such as blended learning and flipped classroom are prevalent in recent years. To implement such learning strategies, learning resources are often put online on learning management systems. The log data on those systems provide an excellent opportunity for getting more understanding about the students through data mining techniques. In this paper, we propose to use sequential pattern mining (SPM) to discover navigational patterns on a learning platform. We attempt to address the lack of literature support about conducting SPM on Moodle. We propose a method to apply SPM that is more appropriate for mining user navigational patterns. We further propose three sequence modeling strategies for mining patterns with educational implications. Results of a study on a statistics course show the effectiveness of the proposed method and the proposed sequence modeling strategies.

Keywords: Sequential pattern mining · Educational data mining · Learning management systems · Moodle · Navigational patterns

1 Introduction

Novel pedagogical approaches supported by digital technologies such as blended learning and flipped classroom are prevalent in recent years. To implement such learning strategies, learning resources such as lecture notes, simulations, videos, and quizzes are often put online on learning management systems (LMSs). Since the LMSs usually store information about students and log their access, they provide an excellent opportunity for getting more understanding about the students through data mining techniques.

Sequential pattern mining (SPM) [6,14] is a data mining technique for finding patterns among sequences of ordered items. It was first proposed for studying the customer purchase sequence for pattern discovery [1]. With the growing interest in e-learning and educational data mining [2,9], SPM has also been applied in education to further facilitate teaching and learning with technology.

Y. Tan et al. (Eds.): DMBD 2017, LNCS 10387, pp. 191–198, 2017.
DOI: 10.1007/978-3-319-61845-6_20

Despite the report of many successful cases showing that the implementation of approaches such as blended learning is effective in enhancing students' learning, few studies have explored the ways that students navigate and interact with the learning resources in the online learning environment and the possible pedagogical implications derived from the activities of students online. One possible reason is that the volume of data acquired for the users' activities can be large. Although current LMSs such as Moodle provide different reports with user statistics, such information might not be sufficient for instructors to draw meaningful conclusion regarding the whole course or behaviors of users online [10,12,13]. In view of this, this study proposes to use SPM to discover sequential patterns, or navigational patterns, on LMSs.

Previous studies have attempted to apply SPM to analyze the usage patterns on various e-learning systems. Surprisingly, few studies have attempted to apply this technique to LMSs for the analysis of learning behavior of students. [15] use both association rule mining and sequential pattern mining to look for resource access patterns of students on Moodle for exam preparation. However, their study does not discuss the potential issues in the implementation of SPM on Moodle. Romero et al. [11] introduce the theoretical and practical way to apply various data mining techniques on Moodle, but the coverage on SPM is limited due to the wide scope of their paper.

This paper attempts to address the lack of literature support about conducting SPM on Moodle. We aim to discover the navigational patterns of students. We propose a method to apply SPM that is more appropriate for mining navigational patterns on LMSs. We further propose three sequence modeling strategies with reference to Zhou et al. [14] for mining patterns with educational implications. We show some results using the proposed method and strategies in a statistics course.

2 Sequential Pattern Mining

Consider a set of items such as resources on a learning management system. A *sequence* is an ordered list of items. It may indicate the order of which resources are accessed by a student. An item can occur multiple times in a sequence. The number of items in a sequence is known as its *length*. A sequence $\alpha = \langle a_1 a_2 \ldots a_n \rangle$ is called a *subsequence* of another sequence $\beta = \langle b_1 b_2 \ldots b_m \rangle$, or α is *contained* in β, if there exists integers $1 \leq j_1 \leq j_2 \leq \cdots \leq j_n \leq m$ such that $a_1 = b_{j_1}, a_2 = b_{j_2}, \ldots, a_n = b_{j_n}$. In other words, the subsequence α can be formed by removing some items from β with the order of the remaining items preserved. For example, the sequence $\langle a, c, d \rangle$ is a subsequence of the sequence $\langle a, b, c, d, e \rangle$, where a, b, c, d, and e are items. Due to brevity, we do not consider the general case where an element of a sequence can be a set of items.

Let $\mathcal{S}$ be a set of sequences. We refer to the number or proportion of sequences in $\mathcal{S}$ that contain a sequence α is known as the *support* of α in $\mathcal{S}$. We denote the support as $\text{support}_{\mathcal{S}}(\alpha)$ and omit the subscript when it is clear from context.

Sequential pattern mining (SPM) aims to discover frequent subsequences among a set of sequences. The frequent subsequences are also called *sequential* *patterns*. Specifically, given a set of sequences $\mathcal{S}$ and a real number $\xi \in [0, 1]$ as threshold, the problem of SPM is to find all the sequences α such that $\text{support}_{\mathcal{S}}(\alpha) \geq \xi$. The threshold ξ is also known as the *minimum support*.

Often we are interested in only the set of closed sequential patterns. A sequential pattern α is *closed* if it is not contained in another sequential pattern that has the same support. More formally, α is closed if there exists no other sequential pattern α' where α is a subsequence of α' and $\text{support}(\alpha) = \text{support}(\alpha')$.

As an example, consider a database with four sequences $\langle a, d, c\rangle$, $\langle a, c, d, c, b, a\rangle$, $\langle b, a, d, c\rangle$, and $\langle a, c\rangle$. The items a, b, c, and d may refer to four different resources or pages on Moodle. If the minimum support is set to be 1.0, the set of sequential patterns resulting from SPM should comprise the three sequences $\langle a\rangle$, $\langle c\rangle$, and $\langle a, c\rangle$. We can check that each of them is contained in all the sequences in the database and hence their supports are all equal to 4. As a counter example, the sequence $\langle a, d\rangle$ has a support of 3 and thus is not included in the result. The result set will be reduced to $\{\langle a, c\rangle\}$ if only closed sequence is considered.

Many algorithms have been developed to mine sequential patterns efficiently. Some of the earliest attempts, e.g. [1], are based on the well-known Apriori algorithm for association rule mining. However, Apriori-like algorithms need to consider an exponential number of candidate sequences in the worst case and they require repeated scanning of the data set to check the support of candidate sequences [6]. Those problems make Apriori-like algorithms infeasible for large data sets or when the sequential patterns are expected to be long.

PrefixSpan [7], on the other hand, avoids those problems by taking another approach. Its main idea is to project the database of sequences into a set of smaller databases based on a set of frequent subsequences. The frequent subsequences are then grown and checked in the smaller projected databases separately. The projection step and the growth step are done recursively until no more longer frequent sequences are found. PrefixSpan has been shown empirically to be considerably faster than another Apriori-like algorithm.

SPM may result in a large number of sequential patterns. This may lead to a long processing time and make the mining results hard to understand and use. Therefore, constraints have been imposed to limit the mining results to those more interesting to users [8]. Among other constraints that have been used, three kinds of constraints are related to our work. The first kind is item constraint, which specifics a subset of items that should or should not be present in the sequential patterns. The second kind is duration constraint, which requires the duration of sequential patterns to be shorter or longer than a given period. The third kind is gap constraint, which requires the time difference between two items in sequential patterns to be shorter or longer than a given gap.

To support the duration and gap constraints, a sequence database has to been extended to contain time information. A time-extended database is defined to be a set of time-extended sequences $\alpha = \langle(t_1, a_1), (t_2, a_2), \ldots, (t_n, a_n)\rangle$, where each item a_i is annotated with a timestamp t_i. The gap between two items with consecutive indices i and j is then defined as $|t_j - t_i|$ and the duration of a sequence is defined as $|t_n - t_1|$.

Algorithms have to been adapted to respect the constraints specified. Hirate and Yamana [5] extend PrefixSpan to support mining sequential patterns with constraints on minimum and maximum gaps and minimum and maximum durations in a time-extended database. Their algorithm was further extended by Fournier-Viger et al. [4] to give only closed sequential patterns.

3 Mining Navigational Patterns on Moodle

Potential Issues. There are some potential issues when conducting sequential pattern analysis based on the log from LMSs. Zhou et al. [14] suggest three main challenges. The primary challenge concerns the granularity level of log data. If fine-grained events such as a single key-stroke or mouse click are recorded by the system, it may obscure the mining of patterns with coarse-grained events. The second challenge concerns the mined results from SPM algorithms. Most existing algorithms are not designed to be applied in educational context, so excessive patterns with limited relevancy and value could be generated, which also cost additional processing time. The last concerns the identification of learning strategies in the pattern analysis process. Domain knowledge would be needed during the mining and post-hoc stage in order to deduce the educational implications from the discovered patterns. In the following, we describe our proposed approach and explain how we handle those three challenges.

Context of Study. We have chosen Moodle[1] to conduct our study for two reasons. First, Moodle is a free and open-source LMS widely used in educational institutions. Second, unlike other cloud-based platforms, Moodle can be deployed on a private server allowing access to the data for data mining.

Our study was conducted in a statistics course. The main types of digital statistics resources provided on Moodle were simulations, online videos and online quizzes in three selected topics: sampling distribution, central limit theorem, and confidence interval. Students could access the learning platform for pre-learning before class. Teachers also offered some learning activities for student to conduct online after the face-to-face lessons. A total of 123 students participated in the course.

Data Extraction. In Moodle, the log data contain fine-grained information about the interaction behaviors of students on the system. Consider the quiz activity on Moodle as an example. Moodle has different descriptions in the log to record every action of students when answering the quizzes (e.g. "view the quiz", attempt the quiz, "submit the quiz", "review the quiz", etc.). If we include all those actions, we may find many trivial patterns showing sequences of work on a single quiz. Therefore, we aggregate those actions by combining consecutive

[1] https://moodle.org/.

Table 1. Targeted actions under study.

Behavior	Description from Moodle as in the pattern
Visited simulation	Viewed the url "[name of the simulation]" e.g. Viewed the url "[Simulation] What is sampling distribution"
Watched online videos	Viewed the page "[name of the video]" e.g. Viewed the page "[Video] What is sampling distribution"
Answered questions within the quizzes	Has viewed the attempt for the quiz "[name of the quiz]" e.g. Has viewed the quiz "[Quiz] Definition of sampling distribution"

quiz-related actions into a single quiz action. We originally recorded the access to individual questions in a quiz, but it also turned out to be excessive.

The actions (or items) we included in the sequential pattern analysis are listed in Table 1. The time at which an action is taken is used as the timestamp in the input to a SPM algorithm if it is needed.

SPM Algorithms. We used a Java library called SPMF [3] for the implementation of SPM algorithms. We consider three algorithms in our study. PrefixSpan is used due to its efficiency. To allow constraints, we use also the algorithm by [4], called Fournier08 below following the convention in SPMF. We use the maximum gap constraint to restrict the time interval between two actions in a sequential pattern to be shorter than 30 min.

An issue was discovered with Fournier08 during our trial. The algorithm considers the time gap as part of an item in a sequence. For example, the sequence $\langle(0, a), (1, b)\rangle$ is contained in $\langle(10, a), (11, b)\rangle$, but not in $\langle(10, a), (20, b)\rangle$ due to difference in time gap. This restriction is unrealistic in our study because it means that the actions of students must be carried out with exactly the same amount of time elapsed for that sequence to be considered as frequent.

To avoid the above issue, we propose to preprocess the timestamps as follows before we use Fournier08. If the time gap of an action with its previous action is smaller than 30 min, we change the timestamp to 0. Otherwise, we keep the original timestamp. We further specify the maximum duration constraint to 30 min. As a result, the proposed method ignores the time gap difference between two sequences and restricts the time gap to be less than 30 min. Note that the change voids the maximum duration constraint and the resulting sequential patterns could last longer than 30 min in the actual sequences.

Sequence Modeling Strategies. The introduction of time constraint in the SPM algorithm could effectively filter patterns which are not educationally meaningful for the interpretation of students' learning behavior online. However, the pattern discovery process could be further stratified for the exploration of specific patterns under different research contexts. Zhou et al. [14] consider this step in processing the log files before pattern discovery as *sequence modeling.* In the current study, three different sequence modeling strategies are proposed. They aim to yield results with various implications in different perspectives.

Quiz-performance sequence modeling. The students in the course are divided into different groups based on their performance in the online quizzes available in the system. The sequential patterns obtained from each group could tell the common patterns leading to good results in the course or the patterns which indicate students are not performing well in the course.

Resource oriented sequence modeling. Sometimes we are interested about the access situation of certain resources. For example the access patterns before or after watching a video. Therefore, we can include only students that have used certain resources in the analysis. The support of a sequential pattern would then refer to the proportion of students following the pattern among those who have used certain resources. An additional advantage of this filtering is the significant reduction of processing time.

Evaluation oriented sequence modeling. The integration of evaluation results based on online questionnaires and log data could provide more information for us to investigate the behavior of students on the platform. Results from questionnaires, such as those adopting a Likert-scale, could only provide an average numerical rating for reference and hence offering limited insights. However, if students are grouped based on the evaluation results (e.g. high and low ratings) before aggregating their log data for pattern discovery, the obtained patterns could further reveal how the ratings from students reflect their navigational behaviors. Teachers could be better informed about the needs of students for possible adjustments on the platform and resources.

4 Empirical Results

In this section we present some results from the study described in the previous section. We first compare the three SPM algorithms described in Fig. 1. As shown in the left chart, PrefixSpan is drastically slower than the other two methods with time constraints. The proposed method is slower than Fournier08 when the minimum support is low. The reason can be explained by the right chart. We see that the proposed method can discover much more sequential patterns with low minimum support. Besides, our experimental results show that when minimum support is set to 0.2, the length of sequential patterns found by Fournier08 is at most one, whereas the length of those found by the proposed method can be seven. This shows that our proposed method can successfully relax the unnecessarily restrictive constraints imposed by Fournier08.

We now look at the sequential patterns returned by the proposed method when minimum support was set to 0.2. Some generated patterns indicate that students often attempted a quiz after watching a video on the same topic. For example, 40% of the students attempted the quiz “Definition of sampling distribution” after watching the video “What is sampling distribution”. On the other hand, there are some patterns containing only quiz activities. For example, there is a pattern with 0.2 support showing the access solely to seven different quizzes,

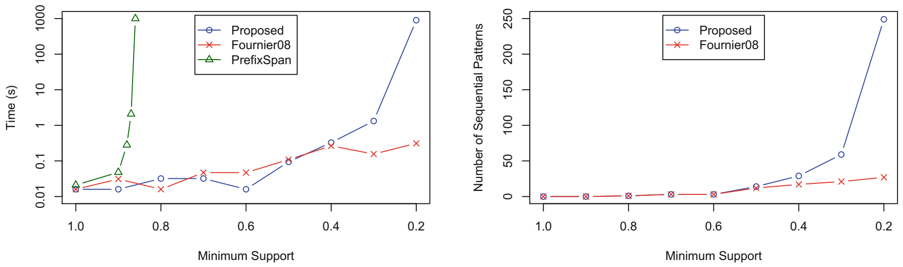

Fig. 1. Comparing the proposed method with FOURNIER08 and PREFIXSPAN.

which implies 20% of students attempted all the quizzes without taking any break longer than 30 min. This might suggest students with such pattern were unengaged in learning online and finished the quizzes in a casual manner, rather than using the quizzes and feedbacks for reflection and knowledge consolidation. These students hence deserve more attention from teachers.

We also tested two sequence modeling strategies in the study. The results show that these strategies can yield some interesting patterns.

For quiz-performance sequence modeling, the overall mean score of all quizzes were calculated and students were divided into two groups, one group with students scored above the mean and another below the mean. SPM was ran separately on both groups' data and the generated patterns were compared. The patterns show that those students scored above the mean demonstrate a better utilization of the learning resources. For example, there are two patterns "viewing a video → attempting a quiz → using statistical simulations → attempting a quiz" and "viewing a video → viewing a video → using statistical simulation" for the topic of sampling distribution with a support value of approximately 0.4. In contrast, the patterns obtained from the group of below mean score mostly consist of one or two actions only with limited implications.

Among the three types of resources, students were found to be less familiar with the simulations. We used resource oriented sequence modeling to focus on students who had used any one of the three available simulations and obtained three sets of sequential patterns. In general, the results show that students who have accessed the simulations would also access other kinds of resource. Among the 28 students who used the simulation for confidence interval, some exhibited navigational behaviors in accessing other resources within the same topic such as "viewing a video → attempting a quiz → viewing a video → attempting a quiz → using the statistical simulation". Such behaviors indicate an effective use of online resources for learning.

5 Conclusion

This paper attempts to provide some insights for researchers who are interested to investigate the navigational patterns of students on Moodle. We discuss some challenges in mining navigational patterns with SPM on Moodle. We propose a method for SPM based on FOURNIER08 [4] and further propose three sequence modeling strategies. Our results show that the proposed method can find more patterns than FOURNIER08 and the proposed strategies can discover patterns that cannot be found without them.

Acknowledgment. The study was funded by Teaching Development Grant (HKIED7/T&L/12-15) under the Hong Kong University Grants Committee.

References

1. Agrawal, R., Srikant, R.: Mining sequential patterns. In: ICDE (1995)
2. ElAtia, S., Ipperciel, D., Zaïane, O.R. (eds.): Data Mining and Learning Analytics: Applications in Educational Research. Wiley, Hoboken (2016)
3. Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.W., Tseng, V.S.: SPMF: a java open-source pattern mining library. J. Mach. Learn. Res. **15**, 3569–3573 (2014)
4. Fournier-Viger, P., Nkambou, R., Nguifo, E.M.: A knowledge discovery framework for learning task models from user interactions in intelligent tutoring systems. In: Gelbukh, A., Morales, E.F. (eds.) MICAI 2008. LNCS (LNAI), vol. 5317, pp. 765–778. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88636-5_72
5. Hirate, Y., Yamana, H.: Generalized sequential pattern mining with item intervals. J. Comput. **1**(3), 51–60 (2006)
6. Mooney, C.H., Roddick, J.F.: Sequential pattern mining - approaches and algorithms. ACM Comput. Surv. **45**(2), 19:1–19:39 (2013)
7. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H.: PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. In: ICDE (2001)
8. Pei, J., Han, J., Wang, W.: Constraint-based sequential pattern mining: the pattern-growth methods. J. Intell. Inf. Syst. **28**, 133–160 (2007)
9. Peña-Ayala, A. (ed.): Educational Data Mining: Applications and Trends. Springer, Cham (2014)
10. Psaromiligkos, Y., Orfanidou, M., Kytagias, C., Zafiri, E.: Mining log data for the analysis of learners' behaviour in web-based learning management systems. Oper. Res. Int. J. **11**(2), 187–200 (2011)
11. Romero, C., Ventura, S., García, E.: Data mining in course management systems: Moodle case study and tutorial. Comput. Educ. **51**(1), 368–384 (2008)
12. Valsamidis, S., Kontogiannis, S., Kazanidis, I., Karakos, A.: E-learning platform usage analysis. Interdiscip. J. E-Learn. Learn. Objects **7**(1), 185–204 (2011)
13. Zaiane, O.R., Luo, J.: Towards evaluating learners' behaviour in a web-based distance learning environment. In: Proceedings of IEEE International Conference on Advanced Learning Technologies, pp. 357–360 (2001)
14. Zhou, M., Xu, Y., Nesbit, J.C., Winne, P.H.: Sequential pattern analysis of learning logs: methodology and applications. In: Handbook of Educational Data Mining, pp. 107–121. CRC Press (2010)
15. Ziebarth, S., Chounta, I.-A., Hoppe, H.U.: Resource access patterns in exam preparation activities. In: Conole, G., Klobučar, T., Rensing, C., Konert, J., Lavoué, É. (eds.) EC-TEL 2015. LNCS, vol. 9307, pp. 497–502. Springer, Cham (2015). doi:10.1007/978-3-319-24258-3_46