An effect-size-based temporal interestingness metric for sequential pattern mining

Yingbin Zhang
University of Illinois at Urbana-Champaign
yingbin2@illinois.edu

Luc Paquette
University of Illinois at Urbana-Champaign
lpaq@illinois.edu

ABSTRACT

Sequential pattern mining is a useful technique for understanding learning behavior. However, it can be challenging to select the most "interesting" patterns discovered through sequence mining. The work presented in this paper proposes an effect-size-based (ESB) method to help researchers identify temporally interesting sequential patterns. ESB is extended from the Temporal Interestingness of Patterns in Sequences (TIPS) technique [4] and distinguishes itself by 1) considering a different association direction between the sequential pattern usage and time, 2) providing a more interpretable ranking metric, and 3) providing a different ranking order for temporally interesting sequential patterns. Both ESB and TIPS are applied to interaction log data to demonstrate their differences in selecting sequential patterns.

Keywords

Sequential pattern mining, effect size, interestingness metric, learning behavior evolution.

1. INTRODUCTION

Sequential pattern mining (SPM) aims to find temporal relationships between events [1]. It is a useful tool to understand students' learning behavior and becomes increasingly popular in the field of education [10, 18]. For example, SPM has been applied to investigate the evolution of cognitive and metacognitive behavior within a computer-based science learning environment [7], to understand students' problem-solving behavior and to explore the associations among metacognitive monitoring, scientific inquiry skills, and task performance within game-based learning environments [4, 16].

Due to the exploratory nature of SPM, researchers need to expend considerable efforts to interpret them and obtain actionable insight for teaching and learning from the discovered sequential patterns [5]. However, the number of sequential patterns discovered through SPM may be huge, and, as such, it is inefficient and sometimes impossible to investigate these patterns one by one. To ease selecting patterns, researchers proposed interestingness metrics to rank sequential patterns or association rules [9].

There has been interest in the topic of temporal analyses of learning data [11], especially in the context of self-regulated learning [12]. As such, patterns that vary across time may be particularly interesting because they may reveal additional information. For instance, the variation of pattern occurrences across time can be used to evaluate the effectiveness of learning support [7]. If the evolution of some sequential patterns in the group who received support from the environment is different from the group without support, the support may have effects on learners' behavior. The evolution of sequential patterns may also provide insights into improving the learning environment. For example, if a sequential pattern beneficial for learning frequently occurs during the whole learning processes except for a particular period, what happens in this period may be interesting. Understanding events in this period may further inform designing intervention to prevent students from stopping this behavior pattern in this period.

In order to ease the selection of temporally interesting patterns, Kinnebrew, Segedy, and Biswas [5] proposed the Temporal Interestingness of Patterns in Sequences (TIPS) technique, an information gain-based approach, to rank patterns contingent on their variation across time. This research extends the TIPS technique by proposing an effect-size-based (ESB) method. ESB was applied to interaction log data of students' using the Betty's Brain learning environment [2] to demonstrate its relative advantages in identifying temporally interesting sequential patterns in comparison with TIPS.

1.1 The procedure of TIPS

TIPS firstly segments each student's log file into n ordered bins (e.g., five ordered bins) with equal sizes [5]. Then, for each student, it calculates the occurrences (also known as instance values) of each frequent sequential pattern in each bin. Thirdly, it takes the occurrences of a pattern per bin per student as the feature and the bin number (e.g., 1, 2, 3, 4, 5) as the label and calculates the information gain (IG) of this pattern. IG refers to the reduction in Shannon entropy about the label from knowing the feature. Its calculation is [14]:

$$IG(L,F) = Entropy(L) - Entropy(L|F) \tag{1}$$

L refers to the label, while F refers to the feature. $Entropy(L)$ is the priori Shannon entropy about the label, while $Entropy(L|F)$ is the conditional Shannon entropy about the label given the feature. Finally, TIPS ranks all frequent sequential patterns based on their IG, and the top-ranking sequential patterns may be temporally interesting.

2. Effect-size based (ESB) temporal interestingness metric

2.1 The procedures of ESB

The ESB approach also needs the first two steps of TIPS, i.e., computing the occurrences per bin per student for each sequential pattern. However, in the next step, the ESB adopts the idea of

Yingbin Zhang and Luc Paquette "An effect-size-based temporal interestingness metric for sequential pattern mining" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 720 - 724

repeated-measures designs [13] and regards the occurrences of a pattern as a variable that is measured several times. One bin is one time. Under this framework, one-way repeated ANOVA can be conducted with the occurrences as the dependent variable and the bin number (i.e., the time) as the independent within-subject variable. Then, ESB calculates the effect size to characterize the association strength between the bin number and the occurrences of a pattern. The ESB regards the effect size as a temporal interestingness metric for sequential patterns. Given that the number of students and bins within a study is constant, the sample sizes for all frequent patterns are the same. Thus, the effect size is comparable across sequential patterns.

There are several effect size measures for ANOVA. Lakens [8] suggested using omega squared for comparisons of effects within a study. The meaning of omega squared is analogous to R squared in linear regression. It estimates the percent of variance explained by the independent variable (the bin in this case). For instance, an omega squared of 0.1 means that 10% of pattern occurrence variance can be explained by the bin number (i.e., time).

Omega squared is used for parametric repeated ANOVA. However, in practice, the distribution of sequential pattern occurrences may violate the assumptions of parametric ANOVA, such as the normality assumption and the homogeneity of variance. For example, some temporally varying sequential patterns may rarely happen at the beginning or end of learning activities. Their occurrence values have many zero in these periods, and their distributions are highly skewed. In this case, it would be better to conduct a non-parametric repeated ANOVA, such as the Friedman test. The effect size corresponding to the Friedman test is Kendall's W [17]. Its calculation is:

$$W = \frac{\chi^2}{N(k-1)} \quad (2)$$

χ^2 is the Friedman test statistic value. N is the number of subjects, and k is the number of measurements per subject. Kendall's W is interpreted similarly to omega squared and ranges from 0 (no relationship) to 1 (perfect relationship).

2.2 Differences between TIPS and ESB

2.2.1 Implicit assumptions.

The direction of the relationship between the occurrences of patterns and time is opposite in the two methods. TIPS examines the extent to which the occurrences of a pattern can distinguish different bins. In other words, TIPS implies that the occurrences of a pattern influence the bin number. In contrast, the ESB assumes that the bin number influences the occurrences of a pattern. While both approaches look at the evolution of the usage of patterns, ESB's assumption is more natural since the assumption is that the bin number is fixed, and the frequency of the pattern is what varies over time. Nevertheless, this distinction between the TIPS and ESB is conceptual and may not have a practical impact.

2.2.2 Interpretability.

The interpretability of ESB may be better than TIPS. As demonstrated above, the meaning of effect size (e.g., omega squared and Kendall's W) is straightforward. Besides, for researchers having experiences with ANOVA, they may already be more familiar with such measures of effect size. This characteristic of ESB can facilitate setting a threshold to filter patterns that may be less temporally interesting. For example, a general rule of thumb on magnitudes of Kendall's W is that W higher than 0.1 but smaller than 0.3 represents a small effect, W no less than 0.3 but less than 0.5 represents a medium effect, and W no less than 0.5 is a large effect [3]. If researchers are only interested in patterns that have at least medium variation across time, they can use 0.3 as the Kendall's W threshold to filter patterns. However, it is more challenging to decide the information gain threshold because the scale of information gains depends on contexts, such as the number of categories (i.e., the number of bins) of the label and the number of distinct values of the feature.

3. Application example

In order to demonstrate the differences of TIPS and ESB in identifying temporally interesting sequential patterns, they were applied to data from a recent study where 88 sixth-grade students learned climate change within Betty's Brain, a computer-based learning environment [2]. Students firstly received a training session on how to use Betty's Brain and used it to study climate change in the next four school days around 45 minutes per day. The action logs of students' working on Betty's Brain were analyzed. The output of TIPS and ESB were compared to investigate the relative advantages of ESB.

3.1 Betty's Brain

In Betty's Brain, students learn about scientific phenomena, such as climate change, by teaching Betty, a virtual pedagogical agent. They teach Betty by adding scientific concepts and directed causal links among the concepts on a blank page. Students can access hypermedia resource pages on relevant scientific concepts and causal relationships. Students can evaluate the causal links by asking Betty to take quizzes. By looking at Betty's correct and incorrect answers, students can identify problems in their understanding.

3.2 Data preprocessing

Firstly, irrelevant actions, such as actions initiated by the system, were removed from the raw action logs [6]. Then, actions were contextualized based on the duration and coherence. Viewing quiz results actions were labeled long vs. short, depending on whether the duration was higher than 3 seconds. Reading page actions were labeled long vs. short, depending on whether the duration was longer than 10 seconds. Long reading pages, adding, revising, and marking links were labeled coherent vs. incoherent, depending on whether these actions were based on prior actions [15]. Finally, the same consecutive actions were collapsed into a single action but labeled multiple. For example, two consecutive short reads were collapsed into an action named multiple short read.

3.3 Applying ESB and TIPS

Traditional sequence mining was applied to the preprocessed dataset to get frequent sequential patterns. The threshold for the support value was 0.5. The maximum gap was 2. This step resulted in 176 frequent sequential patterns. Then, each student's preprocessed log file was segmented into five bins of equal size. For each frequent sequential pattern, its occurrences were calculated per bin per student. Next, Kendall's W and IG of each pattern were computed. These patterns were ranked based on Kendall's W and IG, respectively.

3.4 Results

Some patterns had a high W-based ranking but a comparably lower IG-based ranking or a high IG-based ranking but a comparably lower W-based ranking. Table 1 presents the ranking, Kendall's W, and IG of four such patterns. *View notes → Multiple short read* and *Read short → Multiple incoherent read* were ranked in the top 10 based on Kendall's W, but 33rd and 35th based on IG. In contrast, *Short read → Coherent read* and *Taking a quiz → Prompt → Coherent revision* were ranked in the top 10 based on IG, but 37th and 39th based on Kendall’s W.

Figures 1 to 4 use boxplots to display the occurrences of the four patterns in each bin. The evolutions of *View note → Multiple short read* and *Short read → Multiple incoherent read* was quite similar. Both their usage was more frequent in the first bin than in the others and had little variation among the four last bins. Forty percent of students made *View note → Multiple short read* in the first bin, while less than 16% of students made this pattern in the other bins. Similarly, thirty-four percent of students executed *Short read → Multiple incoherent read* in the first bin, but less than 12.5% of students did so in the others.

By contrast, *Short read → Coherent read* and *Taking a quiz → Prompts → Coherent revision* were less frequent in the first bin than the others. Thirty percent of students executed *Short read → Coherent read* in the first bin, but over 55% of students did so in the others. Twenty-five percent of students made *Taking a quiz → Prompts → Coherent revision* in the first bin, but over 40% of students did so in the others.

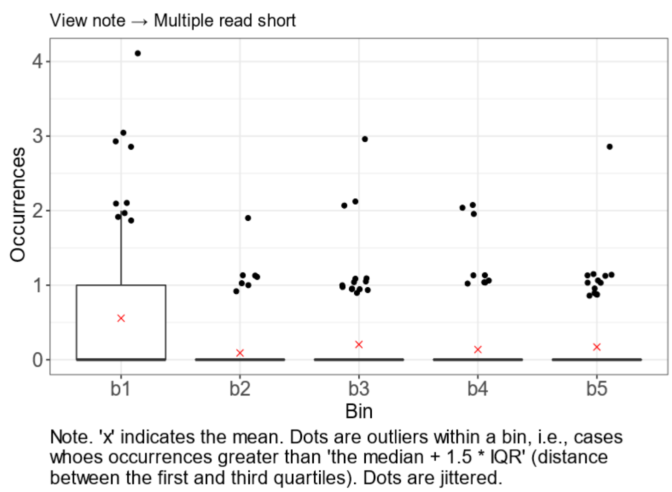

Figure 1. The boxplot of the occurrences of *View note → Multiple short read.*

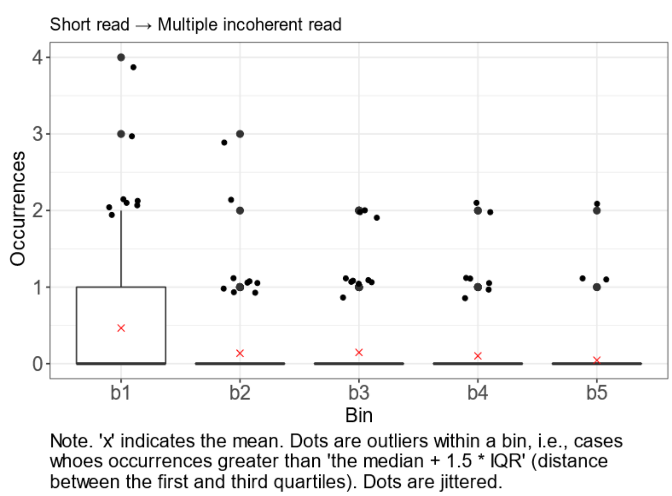

Figure 2. The boxplot of the occurrences of *Short read → Multiple incoherent read.*

There are also similarities between ESB and TIPS. For instance, fourteen patterns were ranked in the top 20 most interesting patterns by both Kendall’s W and IG, and ten patterns were ranked in the lowest 20 by both metrics.

Table 1. Four selected sequential patterns.

Pattern	W - ranking	IG - ranking	Kendall's W	IG
View note → Multiple short read	10	33	0.117	0.051
Read short → Multiple incoherent read	8	35	0.133	0.049
Short read → Coherent read	37	10	0.062	0.073
Taking a quiz → Prompts → Coherent revision	39	7	0.061	0.078

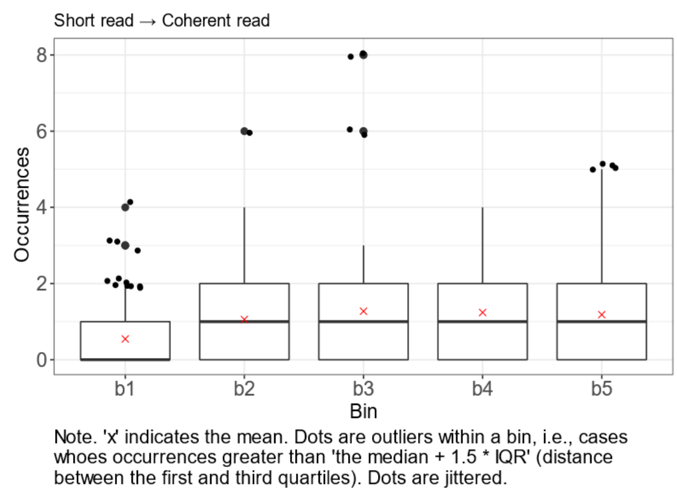

Figure 3. The boxplot of the occurrences of *Short read → Coherent read*.

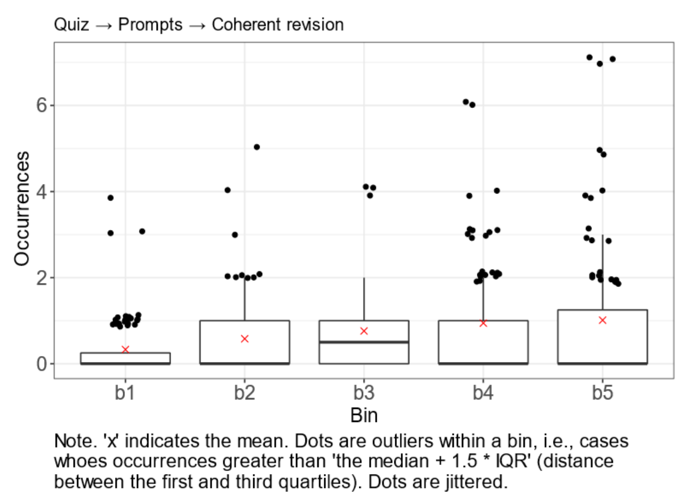

Figure 4. The boxplot of the occurrences of *Taking a quiz → Prompts → Coherent revision*.

4. Discussion

This paper highlighted three differences between ESB and TIPS. The first one is that the implicit assumption of ESB may be more natural than TIPS. ESB assumes that the bin number (i.e., time) influences the occurrences of a pattern, while TIPS implies that the occurrences of a pattern influence the bin number (see section 2.2).

The second difference is the interpretability. It is easier to interpret the ESB metric (i.e., effect size) than the TIPS metric (i.e., IG). For instance, the Kendall's W of *View note → Multiple short read* and *Short read → Multiple incoherent read* was 0.117 and 0.133, respectively, and their IG were 0.051 and 0.049, respectively. A Kendall's W greater than 0.1 but smaller than 0.3 means a small effect [3], so the two patterns have small variation across time. However, it is hard to understand what an IG of 0.051 or 0.049 means as both the number of bins and the number of distinct values of pattern occurrences may influence the range of IG.

The results of the application example revealed the third difference: the rankings of sequential pattern based on the effect size and IG were different. This difference is understandable because the formulas for the effect size and information gain are quite different.

Based on Kendall's W, sequential patterns with more occurrences in the first bin and few occurrences in the others were ranked higher than patterns with fewer occurrences in the first bin and more occurrences in the others. By contrast, based on IG, the former was ranked lower than the latter.

Although for all the above sequential patterns, there is a big difference in pattern usage between the first bin and the others, Kendall's W prefers *View note → Multiple short read* and *Short read → Multiple incoherent read* because the variation of their occurrences across students (between-student variation) were small within each of bin 2 to 5. Recall that less than 16% and 12.5% of students made these patterns in bin 2 to 5, respectively. This means that most of their occurrence values were zero in bin 2 to 5. In contrast, many occurrence values of *Short read → Coherent read* and *Taking a quiz → Prompts → Coherent revision* in bin 2 to 5 was non-zero (over 55% and 40% of students did them, respectively), and their usage had higher variation within each of bin 2 to 5 than *View note → Multiple short read* and *Short read → Multiple incoherent read* (see Figure 1 to 4). In one-way repeated ANOVA, the between-student variation is an error term. Considering the error term, the variation across bins were higher for *View note → Multiple short read* and *Short read → Multiple incoherent read* than for *Short read → Coherent read* and *Taking a quiz → Prompts → Coherent revision*. Therefore, the former patterns had a greater Kendall's W than the latter.

IG prefers *Short read → Coherent read* and *Taking a quiz → Prompts → Coherent revision* because many of their occurrence values in bin 2 to 5 were non-zero, and the distribution of these non-zero values varied across bins. For example, the number of students that did *Short read → Coherent read* two times was the biggest bin 2, but the number of students that did this pattern four times was the biggest in bin 4. Knowing this occurrence differences of *Short read → Coherent read* among bin 2 to 5 could decrease the uncertainty about the bin number (the label). However, most occurrence values of *View note → Multiple short read* and *Short read → Multiple incoherent read* were zero in bin 2 to 5. Knowing their occurrences provided less information about the bin number than knowing the occurrences of *Short read → Coherent read* and *Taking a quiz → Prompts → Coherent revision*.

In summary, the ESB approach may assign higher rankings than TIPS to patterns with more occurrences in one bin but few and similar occurrences in the others, while the latter may assign higher rankings than the former to patterns with fewer occurrences in one bin but more and similar occurrences in the else.

Thus, ESB would be useful if the goal is to identify sequential patterns that mainly appear in only one bin. Such patterns may inform the intervention and learning design. For instance, both *View note → Multiple short read* and *Short read → Multiple incoherent read*, patterns that mainly occurred in the first bin, are generally considered as bad strategies in Betty's Brain. This suggests that students might not be familiar with how to utilize the resource page when they start using Betty's Brain to learn climate change. Therefore, the training session may need to teach students more about how to read the resource page effectively.

4.1 Next steps

The application of TIPS and ESB to the example data provided initial insights about the relative advantages of these approaches, but it is necessary to obtain a more comprehensive understanding of their differences in ranking sequential patterns. This goal will be achieved by conducting a larger scale investigation where TIPS and ESB will be applied to dataset from various learning environments. Such investigation will demonstrate under which situation one method has better utilities than the other so that researchers can make an informed decision about which approach is most appropriate given a research purpose.

While our preliminary application example suggests the utility of ESB to provide insights into improving learning intervention, the goal of the current paper was to propose a new methodological approach for mining temporally interesting sequential patterns. As such, further work will be necessary to leverage ESB to answer formal research questions, such as whether an intervention is effective [7].

5. REFERENCES

[1] Baker, R. 2010. Data mining for education. in McGaw, B., In *International Encyclopedia of Education*, Peterson, P. and Baker, E, Ed. Elsevier Ltd, Oxford, UK, 112-118.

[2] Biswas, G., Segedy, J. R., & Bunchongchit, K. 2016. From design to implementation to practice a learning by teaching system: Betty's brain. *International Journal of Artificial Intelligence in Education*, *26*(1), 350-364.

[3] Field, A. 2013. *Discovering statistics using IBM SPSS statistics*. SAGE, 2013.

[4] Kang, J., Liu, M. and Qu, W. 2017. Using gameplay data to examine learning behavior patterns in a serious game. *Computers in Human Behavior*, *72*. 757-770.

[5] Kinnebrew, J., Mack, D. and Biswas, G., 2013. Mining temporally-interesting learning behavior patterns. In *Proceedings of the 6th International Conference on Educational Data Mining*, (Memphis, TN USA, July 6-9, 2013). EDM '13.

[6] Kinnebrew, J.S., Loretz, K.M. and Biswas, G. 2013. A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining*, *5*(1). 190-219.

[7] Kinnebrew, J.S., Segedy, J.R. and Biswas, G. 2014. Analyzing the temporal evolution of students' behaviors in open-ended learning environments. *Metacognition and Learning*, *9*(2). 187-215.

[8] Lakens, D. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. 863.

[9] Merceron, A. and Yacef, K., 2008. Interestingness measures for association rules in educational data. In *Proceedings of the 1st International Conference on Educational Data Mining*, (Montreal, QC Canada, June 20-21, 2008). EDM '08.

[10] Moon, J. and Liu, Z. 2019. Rich Representations for Analyzing Learning Trajectories: Systematic Review on Sequential Data Analytics in Game-Based Learning Research. In *Data Analytics Approaches in Educational Games and Gamification Systems*, Tlili A., Chang M, Ed. Springer, Singapore, 27-53.

[11] Molenaar, I. (2014). Advances in Temporal Analysis in Learning and Instruction. *Frontline Learning Research*, *2*(4), 15-24

[12] Molenaar, I., & Järvelä, S. (2014). Sequential and temporal characteristics of self and socially regulated learning. *Metacognition and Learning*, *9*(2), 75-85

[13] Myers, J.L., Well, A. and Lorch, R.F. 2010. *Research design and statistical analysis*. Routledge, 2010.

[14] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106.

[15] Segedy, J.R., Kinnebrew, J.S. and Biswas, G. 2015. Using coherence analysis to characterize self-regulated learning behaviours in open-ended learning environments. *Journal of Learning Analytics*, *2*(1). 13-48.

[16] Taub, M. and Azevedo, R. 2018. Using Sequence Mining to Analyze Metacognitive Monitoring and Scientific Inquiry based on Levels of Efficiency and Emotions during Game-Based Learning. *Journal of Educational Data Mining*, *10*(3). 1-26.

[17] Tomczak, M. and Tomczak, E. 2014. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *TRENDS in Sport Sciences*, *21*(1). 19-25.

[18] Van Laer, S. and Elen, J. 2018. Towards a Methodological Framework for Sequence Analysis in the Field of Self-Regulated Learning. *Frontline Learning Research*, *6*(3). 228-249.