

# Mining Temporally-Interesting Learning Behavior Patterns

John S. Kinnebrew  
Department of EECS and ISIS  
Vanderbilt University  
1025 16th Ave S, Ste 102  
Nashville, TN 37212  
john.s.kinnebrew@  
vanderbilt.edu

Daniel L.C. Mack  
Department of EECS and ISIS  
Vanderbilt University  
1025 16th Ave S, Ste 102  
Nashville, TN 37212  
dmack@isis.vanderbilt.edu

Gautam Biswas  
Department of EECS and ISIS  
Vanderbilt University  
1025 16th Ave S, Ste 102  
Nashville, TN 37212  
gautam.biswas@  
vanderbilt.edu

## ABSTRACT

Identifying sequential patterns in learning activity data can be useful for discovering, understanding, and ultimately scaffolding student learning behaviors in computer-based learning environments. Algorithms for mining sequential patterns generally associate some measure of pattern frequency in the data with the relative importance or ranking of the pattern. However, another important aspect of these patterns is the evolution of their usage over the course of a student's learning or problem-solving activities. In order to identify and analyze learning behavior patterns of more interest in terms of both their overall frequency and their evolution over time, we present a data mining technique that combines sequence mining with a novel information-theoretic, temporal-interestingness measure and a corresponding heat map visualization. We demonstrate the utility of this technique through application to student activity data from a recent experiment with the Betty's Brain learning environment and a comparison of our algorithm's pattern rankings with those of an expert. The results support the effectiveness of our approach and suggest further refinements for identification of important behavior patterns in sequential learning activity data.

## Keywords

sequence mining, interestingness measure, information gain, learning behaviors

## 1. INTRODUCTION

Identifying sequential patterns in learning activity data can be useful for discovering, understanding, and ultimately scaffolding student learning behaviors. The primary sequential pattern mining task, as applied in a variety of domains including education, is to discover sequential patterns of items that are found in many of the sequences in a dataset [1]. Some researchers have employed sequential pattern mining to inform student models for customizing learning to individual students (e.g., [2]). Other researchers have employed

sequential pattern mining to better understand learning behavior in particular conditions or groups (e.g., [6, 8]).

However, once these behavior patterns are mined, researchers must interpret and analyze the often large set of patterns to identify a relevant subset of important patterns to investigate or utilize further. Ideally, these patterns provide a basis for generating models and actionable insights about how students learn, solve problems, and interact with the environment. Algorithms for mining sequential patterns generally associate some measure of pattern frequency to rank identified patterns. However, researchers have developed a variety of other measures to utilize properties beyond pattern frequency in ranking mined patterns [4]. These measures are often referred to as "interestingness measures" and have been applied to results from a variety of data mining techniques. To better analyze student learning and behavior, interestingness measures have been used tasks like ranking mined association rules (e.g., [7]).

Investigation of the frequency with which a pattern occurs over time can reveal additional information for pattern interpretation. Further these changes in pattern occurrence may help identify more important patterns, which occur only at certain times or become more/less frequent, rather than patterns with frequent, but uniform, occurrence over time. Qualitatively, we would like to identify patterns that are not rare overall and have significant variations in their frequency over time. In this paper, we present a novel approach, combining sequence mining and an information-theoretic measure for ranking behavior patterns that combines temporal variation in occurrence and overall frequency to provide more effective identification of temporally-interesting patterns. To effectively analyze these patterns and quickly identify trends in the evolution of pattern usage, we employ a related visualization in the form of heat maps. We demonstrate the utility of this technique through application to student activity data from a recent experiment with the Betty's Brain learning environment and a comparison of our algorithm's pattern rankings with those of an expert. The results support the effectiveness of our approach and suggest further refinements for identification of important behavior patterns in sequential learning activity data.

## 2. IDENTIFYING TEMPORALLY-INTERESTING PATTERNS

With long sequences of temporal data, such as student learning activities in a computer-based learning environment, re-

searchers and analysts are not only interested in discovering frequent sequential patterns, but, in many cases, also need to analyze their occurrence over time. In this paper, our focus is on studying how students’ learning behaviors and strategies are employed with differing frequency over the course of learning or problem-solving activities (e.g., as the result of scaffolds and feedback provided by the learning environment or changing demands of the task over the course of learning). In this section, we present the Temporal Interestingness of Patterns in Sequences (TIPS) technique, and corresponding interestingness measure, for identifying and visualizing the most *temporally-interesting* patterns of student behavior.

The first step in analyzing learning activity sequences is to define and extract the actions that make up those sequences from interaction traces logged by the environment. The definition of actions in these sequences for Betty’s Brain data is discussed further in Section 3. Given a set of sequences corresponding to the series of actions performed by each student, the TIPS technique consists of four primary steps:

- 1) Generate candidate patterns that are common to the majority of students by applying sequential pattern mining to students’ learning activity sequences (with a frequency threshold of 50%).

- 2) Calculate a temporal footprint for each candidate pattern by mapping it back to locations where it occurs in the activity sequences. Specifically, each sequence is divided into  $n$  consecutive slices, such that each contains  $\frac{100}{n}\%$  of the student’s actions in the full sequence, where  $n$  is the chosen number of bins defining the temporal granularity of the comparisons. Corresponding slices (e.g., the first slice from each sequence, the second slice from each, and so on) are then grouped into bins and each action in the slices is marked to indicate whether or not it is the beginning of a pattern match in its original sequence. This set of binned and marked actions defines the temporal footprint of the pattern.

- 3) Provide a ranking of the candidate patterns using an information-theoretic interestingness measure (described in more detail below) applied to the temporal footprint of each pattern.

- 4) For the highly-ranked patterns, visualize their temporal footprints using heat maps to more easily assess usage trends and spikes. Specifically, we employ a single-dimensional heat map where each temporal bin’s value is its percentage of the total pattern occurrence. The heat map is generated by assigning a color to each bin, which is determined by where its value falls between the highest and lowest value in the heat map.

In order to identify the more temporally-interesting patterns, the TIPS interestingness measure (in step 3) applies information gain with respect to pattern occurrence across the  $n$  bins of the temporal footprint. Information gain (IG) is defined as the difference in expected information entropy between one state and another state where some additional information is known (e.g., a set of data points considered as a homogeneous group versus one split into multiple groups based on the value of some other feature or attribute). IG is leveraged in classifiers to determine which features are

most discriminatory because they provide the least amount of uncertainty among classes in the data. TIPS applies information gain to determine which patterns are the best descriptors of the data because knowledge of their occurrence provides the least amount of uncertainty about the temporal location of actions in the sequences. In TIPS, IG is applied to the temporal footprint of a pattern by using the  $n$  bins defined in the first step as the classes for the data points, where a data point is a single action in one of the students’ sequences. The feature, in this case, corresponds to whether the action is the start of an occurrence of a given pattern.

For example, when analyzing students’ activity sequences to extract their learning behavior patterns, we may divide up their activity sequence into 5 bins, with each bin containing 20% of the student’s actions in the learning environment. The TIPS measure is then applied for a pattern by determining the information gain between the baseline where approximately the same number of actions are found in any particular bin and the case where we know whether each action corresponded to the occurrence of a pattern, which may happen more or less often in different bins. This IG measure for a pattern defines its temporal-interestingness in TIPS and is used to rank all candidate patterns in descending order, so the pattern that has the highest information gain will be ranked first.

The application of information gain to define the TIPS measure provides two important properties: 1) given two patterns with the same total occurrence, the pattern with the *greater temporal specificity* (i.e., the one that more uniquely distinguishes actions among the periods of time defined by the bins) will have the higher rank, and 2) given two patterns with the same proportions of their total occurrence in corresponding temporal bins, the pattern with the *greater total frequency* will have the higher rank. In this manner, the TIPS measure allows a trade-off between pattern frequency and temporal specificity. Therefore, TIPS tends to emphasize patterns with interesting temporal evolution (e.g., spikes of usage during specific time periods, as well as strongly increasing, decreasing, or peaking trends) even when they are not especially frequent, while also emphasizing particularly frequent patterns with more moderate changes in occurrence over time. Conversely, TIPS tends to deemphasize patterns that are homogeneous over the length of the sequence or that occur rarely.

### 3. BETTY’S BRAIN DATA

The data employed for the analysis in Section 4 consists of student interaction traces from the Betty’s Brain [3]. learning environment. In Betty’s Brain, students read about a science process and teach a virtual agent about it by building a causal map. They are supported in this process by a mentor agent, who provides feedback and support for their learning activities. The data analyzed here was obtained in a recent study with 68 7<sup>th</sup>-grade students taught by the same teacher in a middle Tennessee school. At the beginning of the study, students were introduced to the science topic (global climate change) during regular classroom instruction, provided an overview of causal relations/maps, and given hands-on training with the system. For the next four 60-minute class periods, students taught their agent about climate change and received feedback on both domain con-

**Table 1: Selected Patterns with TIPS and Occurrence Rankings**

Pattern	TIPS Rank	Occurrence Rank	Avg Occurrence
[Read] → [Add link <sup>+</sup> ]	5	12	7.4
[Remove link <sup>-</sup> ] → [Quiz]	22	38	3.8
[Quiz] → [3+ Explanations]	23	78	2.0

tent and learning strategies from the mentor agent.

In Betty’s Brain, the students’ learning and teaching tasks are organized around seven activities: (1) reading resource pages to gain information, (2) adding or removing causal links in the map to organize and teach causal information to Betty, (3) querying Betty to determine her understanding of the domain based on the causal map, (4) having Betty take quizzes that are generated and graded by the mentor to assess her current understanding and the correctness of links in the map, (5) asking Betty for explanations of which links she used to answer questions on the quiz or in queries, (6) taking notes for later reference, and (7) annotating links to keep track of their correctness determined by quizzes and reading. Actions were further distinguished by context details, which for this analysis were the correctness of a link being edited and whether an action involved the same subtopic of the domain as at least one of the previous two actions. The definition of actions in Betty’s Brain learning activity sequences are discussed further in [5].

#### 4. RESULTS

To illustrate and characterize the performance of the TIPS technique, we present selected results of its application to learning activity sequences from the Betty’s Brain classroom study described in Section 3. From the 68 students’ activity sequences, sequential pattern mining identified 215 activity patterns that occurred in at least half of the students. For a broad, initial analysis of their usage evolution over time, we chose to bin pattern occurrence values into fifths of the activity sequences.

Table 1 presents 3 of the top 30 ranked patterns identified by TIPS with their average occurrences per student and a comparison of their rank between TIPS and the baseline ranking by frequency of occurrence. Overall, nearly half of the analyzed TIPS patterns (13 of the top 30) had a rank past 50th by occurrence, with most of those (9 of the top 30) ranking beyond 100th. Such low-ranking (by occurrence) patterns would be unlikely to have been noticed without the TIPS analysis.

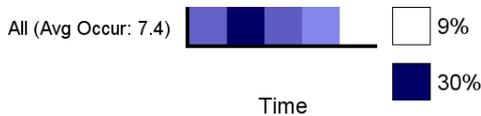
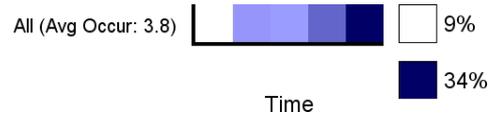
**Figure 1: [Read] → [Add link<sup>+</sup>]**

Figure 1 illustrates the frequency over time for the pattern of reading followed by adding a correct link. This pattern was highly ranked both by occurrence (because it had a high average occurrence) and by TIPS (because it also had strong temporal variation). Students tended to perform this pattern earlier in their learning activities with a peak between

20% and 40% of their complete sequence of activity. An initial estimation of students’ behavior by researchers assumed that students read and added correct links most in the first fifth of their activities with a decreasing trend as the remaining causal relationships were those that were harder to identify. Rather, the identified usage pattern suggests that students require most of an hour working with the system and reading before reaching peak efficiency in determining correct causal links from the resources.

**Figure 2: [Remove link<sup>-</sup>] → [Quiz]**

Another interesting pattern identified by TIPS, which students tended to perform late rather than early, was the removal of an incorrect link followed by taking a quiz. This pattern, illustrated in Figure 2, suggests a monitoring activity in which students employ the quiz to check whether the link was incorrect and should be removed. Although a generally increasing trend was expected since students add more incorrect links over time, the pattern’s occurrence was more heavily weighted toward the end than expected. Over a third of the occurrences were in the last fifth of student activities, suggesting that most students either did not feel the need to monitor their evolving causal map until surprisingly late or took longer than expected to effectively identify potentially incorrect links and/or understand how to use the quiz to verify their removal.

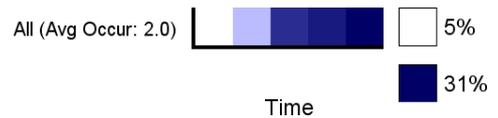
**Figure 3: [Quiz] → [3+ Explanations]**

Figure 3 illustrates another monitoring pattern in which students take a quiz and then ask for explanations of multiple (three or more) quiz questions in a row. This pattern also tends to occur later in the students’ work on the system, but without as clear a peak at the end compared to the previous monitoring pattern. Further, this monitoring pattern had an occurrence rank of 78, making it especially unlikely to have been investigated without TIPS. Although the pattern only occurs twice per student on average, it does suggest that many students attempted to understand and analyze their quiz results in depth as part of their monitoring during the latter half of their work on the system.

To further assess the effectiveness of the TIPS analysis, we performed a preliminary comparison of the TIPS rankings

**Table 2: Expert Interest Ranking**

Measure	High	Medium	Low
TIPS	10	6	4
Occurrence	2	8	10

to those made by an expert - another researcher who coordinated the Betty’s Brain study analyzed here, and who has analyzed student activity data but had no knowledge of the TIPS approach. For this comparison, we used the top 20 patterns from each of the TIPS and occurrence rankings, which resulted in 31 total patterns (9 patterns were in top 20 of both rankings). We presented the expert with only the total occurrence information and the occurrence over time (split into fifths corresponding to how the data was analyzed with TIPS). The order of the patterns was randomized and values for both the overall occurrence and the occurrence over time were represented on separate color-coded scales (between low and high values across all included patterns) to provide some visualization for comparison among the patterns. The expert was asked to group the patterns into three relative categories based on the provided information: high interest (10 patterns), medium interest (10 patterns), and low interest (11 patterns).

Table 2 presents the number of patterns identified by the TIPS and occurrence rankings that the expert grouped into each level of interest. All 10 of the expert’s high interest results were in the top 20 identified by TIPS, with only 2 of them also in the top 20 ranked by occurrence. These results suggest that the TIPS ranking is closer to the expert’s own interest ranking, given the total occurrence and temporal evolution information about each. Next, we presented the expert with the same information but also included the specific activity pattern for each result. When asked to rank the patterns again with this additional information, the results were more equally balanced between the TIPS and occurrence rankings, with six of each in the high interest category, and TIPS having two more than the occurrence ranking in the medium interest category. Overall, these preliminary experiments illustrate the expected point that the activity pattern itself is a major factor in its overall interestingness, but its occurrence and temporal evolution are both important factors. Further, it suggests that rather than relying on only one interestingness measure for identifying potentially important activity patterns, consideration of the top patterns identified by each of multiple measures, including both occurrence and TIPS, may be the most effective way to analyze mined patterns from learning activity sequences.

## 5. CONCLUSION

While identification of common and high-occurrence patterns is undoubtedly useful, finding patterns that have interesting evolution of usage over time is also important for researchers and experts in education, as well as other domains. In this paper, we presented the TIPS technique and interestingness measure for identifying temporally-interesting behavior patterns in learning activity sequences. TIPS is designed to identify patterns with interesting temporal behavior (e.g., spikes of usage during specific time periods or strongly increasing, decreasing, and peaking trends) even when they are not especially frequent, as well as particu-

larly frequent patterns that have at least some clear changes in occurrence over time.

Results from the use of this technique to mine Betty’s Brain data illustrated the potential benefits of identifying behaviors with an interesting evolution over time and helped characterize differences between TIPS and a baseline occurrence ranking. Although general trends in occurrence may be expected for some patterns through consideration of the constraints imposed by the system and the learning activities, TIPS concretely identifies the patterns with strong temporal evolution, confirming some expectations but also identifying patterns with temporal trends that differ from expectations or that would not even have been considered without the TIPS analysis. Further, results from an expert ranking of patterns provided preliminary evidence that patterns identified by TIPS are of particular interest. Overall, the results illustrated the utility of the TIPS technique and suggested that combining the top patterns identified by TIPS and occurrence ranking may be the most useful approach for initial analysis and identification of important learning behavior patterns. Future work will include automatic identification of an effective number of bins for splitting a given set of activity sequences in TIPS and application of identified patterns to improve dynamic scaffolding of learning.

## 6. REFERENCES

- [1] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the Eleventh IEEE International Conference on Data Engineering (ICDE)*, pages 3–14, 1995.
- [2] S. Amershi and C. Conati. Combining unsupervised and supervised classification to build user models for exploratory learning environments. *Journal of Educational Data Mining*, 1(1):18–71, 2009.
- [3] G. Biswas, K. Leelawong, D. Schwartz, N. Vye, and T. Vanderbilt. Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, 19(3):363–392, 2005.
- [4] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9, 2006.
- [5] J. S. Kinnebrew, K. M. Loretz, and G. Biswas. A contextualized, differential sequence mining method to derive students’ learning behavior patterns. *Journal of Educational Data Mining*, In Press, 2013.
- [6] R. Martinez, K. Yacef, J. Kay, A. Al-Qaraghuli, and A. Kharrufa. Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. In *Proceedings of the Fourth International Conference on Educational Data Mining*, Eindhoven, Netherlands, 2011.
- [7] A. Merceron and K. Yacef. Interestingness measures for association rules in educational data. *Educational Data Mining 2008*, page 57, 2008.
- [8] J. Nesbit, M. Zhou, Y. Xu, and P. Winne. Advancing log analysis of student interactions with cognitive tools. *12th Biennial Conference of the European Association for Research on Learning and Instruction (EARLI)*, 2007.