# Mining and Identifying Relationships Among Sequential Patterns in Multi-Feature, Hierarchical Learning Activity Data

Cheng Ye, John S. Kinnebrew, Gautam Biswas
Department of EECS and ISIS
Vanderbilt University
1025 16th Ave S, Ste 102
Nashville, TN 37212
{cheng.ye, john.s.kinnebrew, gautam.biswas}@vanderbilt.edu

***Abstract*—Computer-based learning environments can produce a wealth of information on each student action, which can often be represented at multiple levels of abstraction and with a variety of features. This paper extends an exploratory sequence mining methodology for assessing and comparing students' learning behaviors by autonomously identifying abstraction levels in a hierarchical taxonomy of actions and their potential features. We apply this methodology to action data gathered from the Betty's Brain learning environment. The results illustrate the potential of this methodology in identifying and comparing learning behavior patterns across groups of students with complex, hierarchical action and action feature definitions.**

## I. Introduction

In order to more effectively teach and promote skills required in the modern world, computer-based learning environments (CBLEs) have become more complex and open-ended. In CBLEs, individual student actions can often be represented at multiple levels of abstraction and with a variety of features describing different aspects, contexts, and results of the action.

Sequence mining is widely used in extracting knowledge from databases of human-generated activity data. Further, researchers have applied sequence mining techniques to a variety of educational data in order to better understand and scaffold learning behaviors. In previous work, we have compared sequential patterns derived from student activity sequences to identify ones that differ in usage between two or more groups of students [1], [2] and over time [3].

In this paper, our approach integrates and goes beyond work in differentiating student groups by sequential patterns of behavior [1], [2], as well as work in employing multiple, hierarchically-defined features/dimensions of information in identifying frequent sequential patterns [5], [6]. In particular, the previous work has focused on identifying the *most specific*, detailed frequent sequential patterns, which we extend by identifying a level of specificity (or conversely, generality) that is *most appropriate* for representing sequential patterns that differentiate student groups.

We present example results from the application of this data mining methodology to learning interaction trace data gathered during a middle school class study with the Betty's Brain learning environment. These results illustrate the potential of this methodology in identifying and comparing learning behavior patterns across groups of students with complex, hierarchical action and action feature definitions.

## II. Multi-Feature, Hierarchical, Differential Sequence Mining Methodology

Our approach to effectively mining important patterns in Multi-Feature, Hierarchical (MFH) learning activity sequences employs five primary steps:

1) Define MFH action representation to extract MFH action sequences from student activity traces.
2) Flatten action representation to obtain the most specific action definitions (within frequency constraints) for use with sequence mining methods.
3) Employ DSM to identify differentially-frequent (flattened) activity patterns that distinguish the student groups.
4) Identify hierarchical relationships among mined patterns in the form of directed, acyclic graphs of patterns incorporating different features and levels of detail.
5) Identify the best pattern representations by collapsing more specific pattern nodes into the more general ones that provide a similar degree of differentiation between student groups.

In the final step of this methodology, we iteratively identify and collapse the link for which the parent and child patterns are most similar in terms of their differentiation of the student groups. This similarity is calculated as the difference in effect sizes (by pattern occurrence across the two student groups) between the parent and child patterns with a consideration of the *direction* of the effect (i.e., if the parent pattern occurs more frequently in one student group, but the child parent occurs more frequently in the other student group, then the difference is calculated by summing the effect sizes instead of subtracting them).

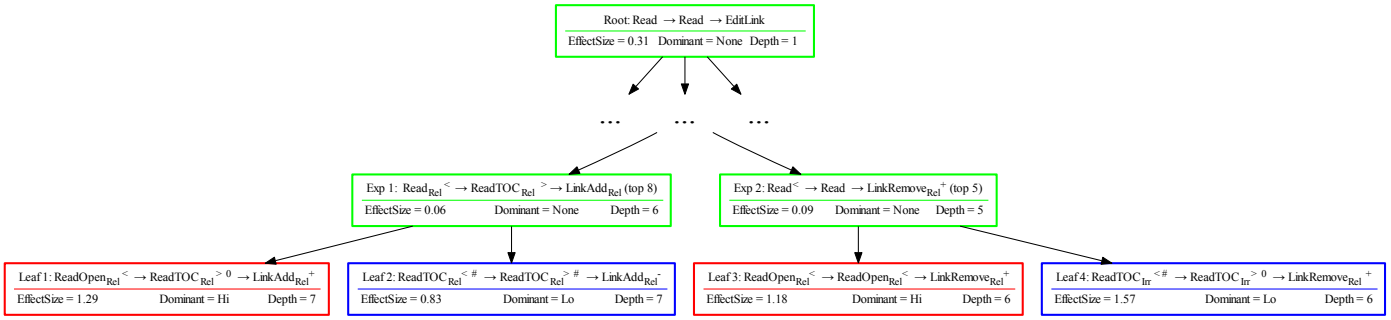


Fig. 1. Pattern tree illustrating some Hi/Lo behavior differences

## III. Results and Conclusion

The data employed for this analysis consists of student interaction traces from the Betty's Brain [4] learning environment. In Betty's Brain, students learn about a science process using a set of hypermedia resources organized into sub-topics by scientific processes and teach a virtual agent, Betty, about what they have learned by building a causal map. In this analysis, we considered the additional action features listed in Table I to analyze the behavior of 8th-grade students from a recent middle Tennessee classroom study in experimental conditions receiving support for identifying causal relationships in the resources. For the differential aspect of the analysis, we focus on the difference between the 16 high-performing (Hi) and 8 low-performing (Lo) students as determined by the quality of their final causal maps.

TABLE I. Action Feature Dimensions

| Actions | Dimension | Value | Symbol |
|---|---|---|---|
| [All except Quiz & Explain] | Relevance | Yes | Rel |
| [All except Quiz & Explain] | Relevance | No | Irr |
| Read | Previous (Full) Read | Yes | ♯ |
| Read | Previous (Full) Read | No | 0 |
| Read | Length | Full | > |
| Read | Length | Short | < |
| EditLink | Map Score Change | Increase | + |
| EditLink | Map Score Change | Decrease | - |
| EditLink | Map Score Change | No Change | = |

With an effect size cutoff of 0.8 to only consider relatively large differences between groups, we identified 312 differential activity patterns, which resulted in 175 pattern trees. In total, there were 913 hierarchical links in the resulting pattern trees and 350 intermediate pattern nodes (i.e., those that are not a leaf pattern identified from the application of DSM nor a root pattern representing the most general form of a set of related leaf patterns). With these pattern trees, we employed the link collapsing described in Section II to identify the most important pattern nodes and hierarchical relationships.

Figure 1 illustrates part of the pattern tree created for a sequence of two reads followed by a map edit, which occurred frequently in both the Hi and Lo group. However, various features of the reading and editing actions allow us to clearly distinguish the Hi group from the Lo group in more specific versions of this pattern illustrated by the lower layer of nodes in Figure 1. In addition to better understanding differences in the skills and approaches, these pattern nodes that are not collapsed into more general versions represent a minimal level of detail that can be used to predict and scaffold students during learning with respect to their likely group characterization.

Conversely, nodes collapsed into their parents represent additional detail that is not particularly important for distinguishing the groups. For link editing followed by taking a quiz and getting an explanation from Betty. Although the initial DSM analysis identified three patterns for link editing followed by taking a quiz and getting an explanation of a quiz question. These leaf pattern nodes, which differed by whether the edit was adding a (correct or incorrect) link or removing an (incorrect) link, were collapsed up to the root pattern early in the search for the best representation level. This indicates that simply following a link edit by a quiz and explanation is characteristic of the Hi group, regardless of the specific details of the link edit, including whether it was correct or not. Thus, this approach to collapsing links in the pattern trees, not only allows the researcher to focus on a smaller subset of important patterns, but also contributes to more accurate interpretation and student characterization during learning by identifying the features and level of specificity necessary for differentiating student groups.

## Acknowledgments

This work was supported by IES CASL grant # R305A120186.

## References

[1] J. S. Kinnebrew and G. Biswas. Identifying learning behaviors by contextualizing differential sequence mining with action features and performance evolution. In *Proceedings of the 5th International Conference on Educational Data Mining (EDM 2012)*, Chania, Greece, June 2012.

[2] 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*, 5(1):190–219, 2013.

[3] J. S. Kinnebrew, D. L. Mack, and G. Biswas. Mining temporally-interesting learning behavior patterns. In S. K. D'Mello, R. A. Calvo, and A. Olney, editors, *Proceedings of the 6th International Conference on Educational Data Mining*, pages 252–255. Memphis, TN, USA, 2013.

[4] K. Leelawong and G. Biswas. Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education*, 18(3):181–208, 2008.

[5] M. Plantevit, A. Laurent, D. Laurent, M. Teisseire, and Y. W. Choong. Mining multidimensional and multilevel sequential patterns. *ACM Transactions on Knowledge Discovery from Data*, 4(1):4:1–4:37, Jan. 2010.

[6] M. Plantevit, A. Laurent, and M. Teisseire. Hype: mining hierarchical sequential patterns. In *Proceedings of the 9th ACM international workshop on Data warehousing and OLAP*, pages 19–26. ACM, 2006.