Sharing and Reusing Data and Analytic Methods with LearnSphere

Ran Liu
ranliu@cmu.edu

Kenneth Koedinger
koedinger@cmu.edu

John Stamper
jstamper@cs.cmu.edu

Philip Pavlik
ppavlik@memphis.edu

ABSTRACT

This workshop will explore LearnSphere, an NSF-funded, community-based repository that facilitates sharing of educational data and analytic methods. The workshop organizers will discuss the unique research benefits that LearnSphere affords. In particular, we will focus on Tigris, a workflow tool within LearnSphere that helps researchers share analytic methods and computational models. Authors of accepted workshop papers will integrate their analytic methods or models into LearnSphere's Tigris in advance of the workshop, and these methods will be made accessible to all workshop attendees. We will learn about these different analytic methods during the workshop and spend hands-on time applying them to a variety of educational datasets available in LearnSphere's DataShop. Finally, we will discuss the bottlenecks that remain, and brainstorm potential solutions, in openly sharing analytic methods through a central infrastructure like LearnSphere. Our ultimate goal is to create the building blocks to allow groups of researchers to integrate their data with other researchers in order to advance the learning sciences as harnessing and sharing big data has done for other fields.

Keywords

Learning metrics; data storage and sharing; data-informed learning theories; modeling; data-informed efforts; scalability.

1. INTRODUCTION

Due to a confluence of a boom of interest both in educational technology and in the use of data to improve student learning, student learning activities and progress are increasingly being tracked and stored. There is a large variety in the kinds, density, and volume of such data and to the analytic and adaptive learning methods that take advantage of it. Data can range from simple (e.g., clicks on menu items or structured symbolic expressions) to complex and harder-to-interpret (e.g., free-form essays, discussion board dialogues, or affect sensor information). Another dimension of variation is the time scale in which observations of student behavior occur: click actions are observed within seconds in fluency-oriented math games or in vocabulary practice, problem-solving steps are observed every 20 seconds or so in modeling tool interfaces (e.g., spreadsheets, graphers, computer algebra) in intelligent tutoring systems for math and science, answers to comprehension-monitoring questions are given and learning resource choices are made every 15 minutes or so in massive open online courses (MOOCs), lesson completion is observed across days in learning management systems, chapter/unit test results are collected after weeks, end-of-course completion and exam scores are collected after many months, degree completion occurs across years, and long-term human goals like landing a job and achieving a good income occur across lifetimes. Different paradigms of data-driven education research differ both in the types of data they tend to use and in the time scale in which that data is collected. In fact, relative isolation within disciplinary silos is arguably fostered and fed by differences in the types and time scale of data used [4, 5].

Thus, there is a broad need for an overarching data infrastructure to not only support sharing and use within the student data (e.g., clickstream, MOOC, discourse, affect) but to also support investigations that bridge across them. This will enable the research community to understand how and when long-term learning outcomes emerge as a causal consequence of real-time student interactions within the complex set of instructional options available [2]. Such an infrastructure will support novel, transformative, and multidisciplinary approaches to the use of data to create actionable knowledge to improve learning environments for STEM and other areas in the medium term and will revolutionize learning in the longer term.

LearnSphere transforms scientific discovery and innovation in education through a scalable data infrastructure designed to enable educators, learning scientists, and researchers to easily collaborate over shared data using the latest tools and technologies. LearnSphere.org provides a hub that integrates across existing data silos implemented at different universities, including educational technology "click stream" data in CMU's DataShop, massive online course data in Stanford's DataStage and analytics in MIT's MOOCdb, and educational language and discourse data in CMU's new DiscourseDB. LearnSphere integrates these DIBBs in two key ways: 1) with a web-based portal that points to these and other learning analytic resources and 2) with a web-based workflow authoring and sharing tool called Tigris. A major goal is to make it easier for researchers, course developers, and instructors to engage in learning analytics and educational data mining without programming skills.

2. SPECIFIC WORKSHOP OBJECTIVES

Broadly, this workshop offers those in the EDM community an exposure to LearnSphere as a community-based infrastructure for educational data and analysis tools. In opening lectures, the organizers will discuss the way LearnSphere connects data silos across universities and its unique capabilities for sharing data, models, analysis workflows, and visualizations while maintaining confidentiality.

More specifically, we propose to focus on attracting, integrating, and discussing researcher contributions to Tigris, the web-based workflow authoring and sharing tool. The goal of Tigris is to support any custom analysis method that can be applied to the datasets and to produce outputs in a standardized way that facilitates both quantitative and qualitative model comparisons. This workflow feature allows researchers to apply their own analysis methods to the vast array of datasets available in the educational data repository. It affords researchers the advantages of (1) using the built-in learning curve visualizations on the outputs of their own analysis workflows, (2) easily comparing their results both quantitatively and graphically to the outputs of

any other analysis methods that are currently in LearnSphere (e.g., Bayesian Knowledge Tracing [1], Performance Factors Analysis [6], MOOC activity analysis [3], and others) or that have been uploaded to LearnSphere as a custom workflow, and (3) sharing their own analysis workflows with the community of researchers. Without any prior programming experience, researchers can use LearnSphere's drag-and-drop interface to compare, across alternative analysis methods and across many different datasets, model fit metrics like AIC, BIC, and cross validation as well as parameter estimates themselves.

Workshop submissions will involve a brief description of an analysis pipeline relevant to modeling educational data as well as accompanying code. Prior to the workshop itself, the organizers will coordinate with authors of accepted submissions to integrate their code into Tigris. A significant portion of the workshop will be dedicated to hands-on exploration of custom workflows and workflow modules within Tigris. Authors of accepted submissions will present their analysis pipelines, and everyone attending the workshop will be able to access those analysis pipelines within Tigris to a variety of freely available educational datasets available from LearnSphere. The end goal is to generate, for each workflow component contribution in the workshop, a publishable workshop paper that describes the outcomes of openly sharing the analysis with the research community.

Finally, workshop attendees will discuss bottlenecks that remain toward our goal of an easier, more open way to share analytic tools. We will also brainstorm possible solutions. Our goal is to create the building blocks to allow groups of researchers to integrate their data with other researchers we can advance the learning sciences as harnessing and sharing big data and analytics has done for other fields.

3. REFERENCES

[1] Corbett, A.T., & Anderson, J.R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

[2] Koedinger, K.R., Booth, J.L., & Klahr, D. (2013). Instructional complexity and the science to constrain it. Science, 342(6161), 935-937.

[3] Koedinger, K.R., Kim, J., Jia, J.Z., McLaughlin, E.A., & Bier, N.L. (2015). Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the 2nd ACM Conference on Learning@ Scale*, pp. 111-120.

[4] Koedinger, K.R., Corbett, A.T., & Perfetti, C. (2012). The Knowledge‐Learning‐Instruction framework: Bridging the science‐practice chasm to enhance robust student learning. *Cognitive science, 36*(5), 757-798.

[5] Newell, A. (1990). Unified theories of cognition. Cambridge, MA: Harvard University Press.

[6] Pavlik, P.I., Cen, H., & Koedinger, K.R. (2009). Performance factors analysis – A new alternative to knowledge tracing. In *Proceedings of the 14th International Conference on AIED*, 531–538.