Workshops

Graph-based Educational Data Mining (G-EDM 2017)

Collin F. Lynch, Tiffany Barnes, Linting Xue, & Niki Gitinabard
North Carolina State University, Raleigh, North Carolina, USA
cflynch, tmbarnes, lxue3, & ngitina@ncsu.edu

ABSTRACT

With the growing popularity of MOOCs and computer-aided learning systems, as well as the growth of social networks in education, we have begun to collect increasingly large amounts of educational graph data. This graph data includes complex user-system interaction logs, student-produced graphical representations, and conceptual hierarchies that large amounts of graph data have. There is abundant pedagogical information beneath these graph datasets. As a result, graph data mining techniques such as graph grammar induction, path analysis, and prerequisite relationship prediction has become increasingly important. Also, graphical model techniques (e.g. Hidden Markov Models or probabilistic graphical models) has become more and more important to analyze educational data.

While educational graph data and data analysis based on graphical models has grown increasingly common, it's necessary to build a strong community for educational graph researchers. This workshop will provide such a forum for interested researchers to discuss ongoing work, share common graph mining problems, and identify technique challenges. Researchers are encouraged to discuss prior analyses of graph data and educational data analyses based on graphical models. We also welcome discussions of in-progress work from researchers seeking to identify suitable sources of data or appropriate analytical tools.

1. PRIOR WORKSHOPS

So far, we have successfully held two international workshops on Graph-based Educational Data-Mining. The first one was held in London, co-located with EDM 2014. It featured 12 publications of which 6 were full-papers, the remainder short papers. Having roughly 25 full-day attendees and additional drop-ins, it led to a number of individual connections between researchers and the formation of an e-mail list for group discussion. The second one was co-located with EDM 2015 in Spain. 10 authors presented their published work including 4 full papers and 6 short papers there.

2. OVERVIEW AND RELEVANCE

Graph-based data mining and educational data analysis based on graphical models have become emerging disciplines in EDM. Large-scale graph data, such as social network data, complex user-system interaction logs, student-produced graphical representations, and conceptual hierarchies, carries multiple levels of pedagogical information. Exploring such data can help to answer a range of critical questions such as:

- For social network data from MOOCs, online forums, and user-system interaction logs:
  - What social networks can foster or hinder learning?
  - Do users of online learning tools behave as we expect them to?
  - How does the interaction graph evolve over time?
  - What data we can use to define relationship graphs?
  - What path(s) do high-performing students take through online materials?
  - What is the impact of teacher-interaction on students' observed behavior?
  - Can we identify students who are particularly helpful in a course?
- For computer-aided learning (writing, programming, etc.):
  - What substructures are commonly found in student-produced diagrams?
  - Can we use prior student data to identify students' solution plan, if any?
  - Can we automatically induce empirically-valid graph rules from prior student data and use induced graph rules to support automated grading systems?

Graphical model techniques, such as Bayesian Network, Markov Random Field, and Conditional Random Field, have been widely used in EDM for student modeling, decision making, and knowledge tracing. Utilizing these approaches can help to:

- Learn students' behavioral patterns.
- Predict students' behaviors and learning outcomes.

- Induce pedagogical strategies for computer-aided learning systems.
- Identify the difficult level of the knowledge components in the intelligent tutoring systems.

Researches related to these questions can help us to better understand students' learning status, and improve the teaching effectiveness and student learning. Our goal in this workshop is to bring together researchers with special interest in graph-based data analysis to 1) discuss state of the art tools and technologies, 2) identify common problems and challenges, and 3) foster a community of researchers for further collaboration. We will consider the submission of full and short papers as well as posters and demonstrations covering a range of graphics topics that include, but are not limited to:

- Social network data
- Graphical solution representations
- Graphical behavior models
- Graph-based log analysis
- Large network datasets
- Novel graph-based machine learning methods
- Novel graph analysis techniques
- Relevant analytical tools and standard problems
- Issues with graph models
- Tools and technologies for graph grammar (pattern) recognition
- Tools and technologies for automatic concept hierarchy extraction
- Computer-aided learning system development involved with graphical representations
- Use of graphical models in educational data

We particularly welcome submissions of in-progress work both from students and researchers with problems who are seeking appropriate analytical tools, and developers of graph analysis tools who are seeking new challenges.

3. WORKSHOP ORGANIZERS

Dr.Collin F. Lynch is an Assistant Professor in the Department of Computer Science at North Carolina State University. His primary research is focused on graph-based educational data mining, the development of robust intelligent tutoring systems, and adaptive educational systems for ill-defined domains such as scientific writing, law, and engineering. In his more recent work he has also been involved in the development of Intelligent Tutoring Systems for Logic and Probability and social networking analysis for research communities.

Dr.Tiffany Barnes is an Associate Professor of Computer Science at NC State University. She received an NSF-CAREER Award for her novel work in using data to add intelligence to STEM learning environments. That grant supported the development of InVis a novel tool that use graph-based representations of student-tutor interaction data to evaluate the impact of intelligent tutoring systems on student problem-solvers and to automatically extract hints and student advice from log data using graph-analysis. More recently she has received grants for the analysis of large-scale online courses and the development of procedural guidance from intelligent tutoring system data.

Linting Xue is a third year Ph.D. student in the Department of Computer Science at North Carolina State University. She is interested in the graph data mining methods for educational graph data. Her current research is focused on automatically graph grammars induction for student-produced argument diagrams. The induced graph grammars can be used as features for automatic grading and provide the hints for argumentative writing.

Niki Gitinabard is a second year Ph.D. student in the Department of Computer Science at North Carolina State University. She is interested in social network analysis in learning environments. She is currently working on social graph generation and analysis based on students' explicit and implicit interactions.

4. WORKSHOP ORGANIZATION

We will organize this workshop as a full or half-day mini-conference with time set aside for paper presentations, large-group discussion, and individual networking. We will open the workshop with a summary of prior meetings. We will spend the morning on presentations with a short discussion session before lunch. The afternoon session will be divided between presentations and working groups which will focus on identifying shared problems, small-group networking, and planning for follow up work. We will invite submissions of full papers which describe mature work. We will also accept short papers describing in-progress work or student projects, and poster/demo submissions for those presenting available data, tools, and methods. This last category is particularly targeted at researchers who have data or methods available and are seeking to identify potential collaborators.

Workshop on deep learning with educational data

Ryan Baker
University of Pennsylvania
Philadelphia, PA 19104
ryanshaunbaker@gmail.com

Joseph E. Beck
Worcester Polytechnic Institute
Worcester, MA, 01609
josephbeck@wpi.edu

Min Chi
North Carolina State University
Raleigh, NC 27695
mchi@ncsu.edu

Neil T. Heffernan
Worcester Polytechnic Institute
Worcester, MA, 01609
nth@wpi.edu

Mike Mozer
University of Colorado Boulder
Boulder, CO 80309
mozer@colorado.edu

1. WORKSHOP TOPIC

This workshop focuses on applications of deep learning for educational data. Deep learning is a machine learning approach using neural networks with multiple levels of representational transformation (i.e., hidden layers). Deep learning has been used in a variety of domains over the past five years with impressive results. Recently, it has been used for educational data sets with mixed results when compared to traditional modeling methodologies.

We are interested in work on a variety of topics with deep learning: new prediction and modeling problems, best practices for featurizing data, network architectures, approaches to pre-training and whether it is necessary, interpreting the learned models, end-to-end deep learning approaches with low-level non-symbolic data, toolkits people have developed, empirical results on known problems to help the field develop best practices. The workshop is also interested in negative results such as analyses of data sets and domains where deep learning fails to achieve state of the art performance.

2. GOALS OF WORKSHOP

The primary goal of this workshop is to provide a venue for researchers to present emerging work. There is not much prior art on applying deep learning to educational data, and it is unclear even what the scope of possible applications are: although most work has focused on student modeling, some work has focused on using deep learning to assist in scoring essays. Having a discussion about possible application areas will be productive.

In addition, this workshop will focus on recent big topics in deep learning for educational data. A paper published in 2016 “How deep is knowledge tracing” questions the need for deep models, and will be discussed at the workshop.

Finally, this workshop will provide researchers on deep learning for EDM a chance to get focused feedback on their work. Ensuring that the research is critiqued by a roomful of people interested in the topic is more useful to the presenters (and the community) than counting on haphazard interactions at the conference.

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.