Panel: Educational Data Mining meets Learning Analytics

Ryan S.J.d. Baker
Dept. Social Science and Policy Studies
Worcester Polytechnic Institute
Worcester, MA USA
+1 508-831-5355
rsbaker@wpi.edu

Erik Duval
Dept. Computerwetenschappen
Katholieke Universiteit Leuven
Celestijnenlaan 200A - bus 2402
3001 Heverlee
België
+32 1632-7700
erik.duval@cs.kuleuven.be

John Stamper
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213, USA
+1 412-268-9690
john@stamper.org

David Wiley
Dept. Instructional Psychology & Technology
Brigham Young University
Provo, UT 84602, USA
+1 801-422-7071
david.wiley@byu.edu

Simon Buckingham Shum (Panel Chair)
Knowledge Media Institute
The Open University
Walton Hall
Milton Keynes, MK7 6AA, UK
+44-1908-655723
s.buckingham.shum@gmail.com

ABSTRACT

This panel continues the dialogue between the *Educational Data Mining* and *Learning Analytics* communities. *EDM* has been developing as a community for longer than the *LAK* conference, so what if anything makes the LAK community different, and where is the common ground? Is LAK just reinventing the wheel, or adding some important new spokes? To push the metaphor, are LAK's wheels fit for the new learning terrain without EDM? In any case, what do we need in addition to wheels to go places? Is EDM "narrower but deeper", best suited for stable, well understood domains in which domain knowledge and user cognition can be formally modelled, but at considerable expense? Is EDM also more mathematical, while LA is more qualitative, socially oriented, and interested in open, social learning "in the wild" where far less can be known about users or learning objectives? Or are these just myths and stereotypes waiting to be debunked? Two representatives from each community (LAK: Duval & Wiley; EDM: Baker & Stamper) will present a brief position, outlined in this paper, in which they set out what it is that excites them about their 'home' discipline and community, and how they see the relationships between the fields. The issue will then be opened up for conference delegates to debate what could or should be future trajectories for the fields.

Categories and Subject Descriptors

J.1 [**Administrative Data Processing**] *Education*; K.3.1 [**Computer Uses in Education**] *Collaborative learning, Computer-assisted instruction (CAI)*

General Terms

Design, Human Factors, Theory

Keywords

learning analytics; educational data mining

LAK12: 2nd International Conference on Learning Analytics & Knowledge, 29 April – 2 May 2012, Vancouver, BC, Canada

1. EDUCATIONAL DATA MINING

1.1 Ryan Baker

I will discuss a vision of what the educational data mining community has to offer to the science and practice of education, focusing on the positive things that LAK can learn from EDM, and the positive things that EDM can learn from LAK. As such, I believe that the future of EDM and LAK should be as best friends, rather than as frenemies (or worse), as is so often seen when two research communities occupy similar spaces in their scope of scientific inquiry.

My belief is that one of the key contributions that EDM makes is the advancement of rigorous positions on how to verify that models produced through data mining and analytics are valid and generalizable. The migration of some of these standards and approaches to the LAK community may be useful to LAK researchers in specific cases.

In addition, the EDM community's focus on comparing different modeling methods, towards discovering when specific models and frameworks are appropriate, is producing knowledge that would be beneficial to researchers in the LAK community.

At the same time, I believe that many of the problems being attacked in LAK and the methods for attacking these problems are unique and highly advanced, and that EDM would benefit from

learning more about these problems and approaches. I also believe that LAK's attention to the needs and concerns of various stakeholder groups is exemplary, and that EDM research would benefit strongly from learning how LAK researchers are addressing those issues.

Furthermore, in some areas EDM research and LAK research are making advances that could be integrated, to the benefit of both communities. For example, much EDM research now leverages human judgment to support data mining, suggesting that the combination of LAK research in leveraging human judgment with EDM expertise in data mining may produce positive results beyond the current capacities of either community.

Finally, I will argue vigorously (but in a friendly way) against the proposition that EDM is narrower but deeper, best suited for stable, well understood domains. In my opinion, both EDM and LAK are highly suited for any learning domain, and exemplary research in both communities is targeted at extending the phenomena and settings in which learning and learners can be studied.

Bio: Ryan S.J.d. Baker is Assistant Professor of Psychology and the Learning Sciences at Worcester Polytechnic Institute. His research, at the intersection of educational data mining/learning analytics and human-computer interaction, focuses on modeling and studying students' learning, engagement, and affect. He was elected founding President of the International Educational Data Mining Society in 2011, and is Associate Editor of the Journal of Educational Data Mining. He graduated from Carnegie Mellon University in 2005, with a Ph.D. in Human-Computer Interaction. He received the Best Paper Award at the Intelligent Tutoring Systems Conference in 2006, and received the Best Oral Presentation Award at the Intelligent Tutoring Systems Conference in 2010. http://users.wpi.edu/~rsbaker

1.2 John Stamper

In the past, I was always preaching: "the data flood is coming," but today that has changed to: "the data flood is here!" Traditional methods of data analysis have not kept pace with the amount of data that can be collected and is being collected from educational environments today. Many others have also seen this trend which is one of the main reasons that the Educational Data Mining and Learning Analytics communities have begun to grow as fast as they have in the last couple of years.

One of my roles is the Technical Director of the Pittsburgh Science of Learning Center DataShop[1], which has become a large repository of log data collected from a variety of educational systems, most notably the cognitive tutors that have been developed at Carnegie Mellon University and Carnegie Learning, Inc. The datasets in DataShop are composed of fine-grained data, with student actions recorded roughly every 20 seconds. As of March 2012, DataShop contains over 300 datasets which are comprised of over 70 million student actions and 190,000 student hours of data. Over time, we have seen a shift in the types of data collected. Originally, most of the datasets were from experimental studies performed in a classroom and generally lasted days or weeks. More often now, the data coming in has a much longer time frame lasting months, semesters, or entire years. In addition to study data, we are now receiving course data that does not represent any preset experiment but is collected in hopes that researchers can use the data to understand learning and improve the courses where the data was derived.

In 2010, DataShop hosted the KDD Cup Challenge, which asked participants to predict student performance on mathematical problems from logs of student interaction data similar to the type stored in DataShop. One major difference was that the size of the datasets for the competition was larger than the entire DataShop repository at that time. The size did seem to provide major hurdles for researchers in the competition – especially those from the learning sciences. To effectively use these large datasets to make discoveries, both the EDM and LAK communities need to develop or find the tools and algorithms to handle the size and robustness of these data.

For the most part, the goals of the EDM and LAK communities overlap extensively, but there are subtle differences that I see between the two communities. The most fundamental is where the research is focused. The EDM community tends to use data to understand how and when learning occurs. The focus is on the process. One key area is building predictive models to explain and detect aspects of learning. The LAK community tends to focus on the learner, and using data to explore how the learner interaction with technology affects individual learning. Again, the difference is subtle, and both are needed to improve the effectiveness of educational technology, which is the goal of both communities.

Bio: John Stamper is a member of the research faculty at the Human-Computer Interaction Institute at Carnegie Mellon University. He is also the Technical Director of the Pittsburgh Science of Learning Center DataShop. His primary areas of research include Educational Data Mining and Intelligent Tutoring Systems. As Technical Director, John oversees the DataShop, which is an open data repository and set of associated visualization and analysis tools for researchers in the learning sciences. John received his PhD in Information Technology from the University of North Carolina at Charlotte, holds an MBA from the University of Cincinnati, and a BS in Systems Analysis from Miami University. Prior to returning to academia, John spent over ten years in the software industry including working with several start-ups. He is a Microsoft Certified Systems Engineer (MCSE) and a Microsoft Certified Database Administrator (MCDBA). John was the co-chair of the 2010 KDD Cup Competition, titled "Educational Data Mining Challenge," which centered on improving assessment of student learning via data mining. http://www.hcii.cmu.edu/people/faculty/john-stamper

2. LEARNING ANALYTICS

2.1 Erik Duval

In my view, Learning Analytics is about *collecting traces that learners leave behind and using those traces to improve learning*. Educational Data Mining can process the traces algorithmically and point out patterns or compute indicators. My personal interest is more in using the traces in order to empower learners to be 'better learners'.

My team focuses on building dashboards that visualize the traces in ways that help learners or teachers to steer the learning process. I like this approach because it focuses on helping people rather than on automating the process. It is inspired by a 'modest computing' approach[2] where the technology is used to support what we want people to be good at (being aware of what is going

[1] https://pslcdatashop.web.cmu.edu/about

[2] http://www.teleurope.eu/pg/podcasts/play?g=140221

[2] http://www.teleurope.eu/pg/podcasts/play?g=140221

on, making decisions, …) by leveraging what computers are good at (repetitive, boring tasks…).

Of course, capturing *meaningful* learning traces is something that both we and the EDM community struggle with. Translating those traces into *visual representations and feedback* that support learning is another challenge: the danger of presenting meaningless eye candy or networks that confuse rather than help is all too real.

Both our work and that of the EDM community is also difficult to *evaluate*: we can (and do!) evaluate usability and usefulness, but assessing real learning impact is hard – both on a practical, logistical level (as it requires longitudinal studies) as well as on a more methodological level (as impact is 'messy' and it is difficult to isolate the effect of the intervention that we want to evaluate).

In both these areas, we may be able to make better progress by exchanging our experiences. There is also an opportunity to combine both approaches: for instance, we can use visualization techniques to help people understand what data mining algorithms come up with and why. In that way, work on visualization can help to increase understanding of and trust in what the EDM community achieves.

Bio: Erik is professor of computer science and chairs the research unit on human-computer interaction, at KU Leuven, the University of Leuven in Belgium. His research focuses on novel ways to interact with information, through information visualization, mobile information devices and multi-touch displays. Typical application areas are technology enhanced learning, interaction with music and 'research2.0′. Erik teaches courses on Human-Computer Interaction, Multimedia, problem solving and design. He is a member of the informatics section of the Academia Europeae and co-founded two spin-offs on personalized smart interaction with music and scientific output, as well as the not-for-profit ARIADNE Foundation that promotes share and reuse of learning material. http://erikduval.wordpress.com/about

2.2 David Wiley

As part of his 2 sigma work, Bloom (1984)[3] challenged educational researchers to devise *practical methods* – "methods that the average teacher or school faculty can learn in a brief period of time and use with little more cost or time than conventional instruction" – that would help learners reach their academic potential. My personal interest in learning analytics lies in its ability to answer extremely practical and socially responsive questions such as, "What is the most effective thing a teacher could do with her next 30 minutes?" and "What is the most effective experience a learner could choose next?"

In my view, learning analytics as a term simply describes the extremely interdisciplinary endeavor of providing this pragmatic support for learning.

On the "back end" learning analytics combines knowledge and techniques from data mining and psychometrics to leverage both behavioral data and data about academic performance. From this perspective learning analytics is a synthesis of techniques like Naïve Bayes, Rasch modeling, collaborative filtering, and item response theory. Both data mining and psychometrics possess a rich set of tools that are applicable to the problems we want to solve using learning analytics.

On the "front end" learning analytics combines knowledge and techniques from data visualization and UI/UX to empower ordinary teachers or learners with little or no training to bring the full power of data to bear on their learning-related decisions. Data-related tools still look too much like the "Your Product" in the famous StuffThatHappens comic.[4] We typically fail to acknowledge that the work involved in achieving Google or Apple-like simplicity in the front end design of learning analytics tools will require at least as much effort and attention as will solving "back end" problems.

Learning analytics, then, is a consumer of the knowledge created by the educational data mining community and depends on this and the work of numerous other fields in order to bring the full promise of technology (in this case, the data-enabled promises) to ordinary learners and teachers everywhere.

Bio: Dr. David Wiley is Associate Professor of Instructional Psychology and Technology, and Associate Director of the Center for the Improvement of Teacher Education and Schooling at Brigham Young University, where he directs the Open Education Group. David is currently Senior Fellow for Open Education at the National Center for Research in Advanced Information and Digital Technologies (Digital Promise) and a Peery Social Entrepreneurship Research Fellow in BYU's Marriott School of Business. Previously, David was a recipient of the National Science Foundation's CAREER grant. http://opencontent.org/blog

[3] http://en.wikipedia.org/wiki/Bloom's_2_Sigma_Problem

[4] http://devio.wordpress.com/2012/02/19/user-interface-design/