Data sources for learning analytics: What is the current state-of-art in China?

Visiting scholar report to Shanghai Open University, July 2017

Tore Hoel
Oslo and Akershus University of Applied Sciences, Norway
tore.hoel@hioa.no

Abstract
This short report on research conducted during a visiting scholarship at Shanghai Open University, June 2017, explores the current state-of-affairs in learning analytics, also known as educational big data, in China with a focus on which data sources are used for analysis as documented by recent Chinese scholarly publications. The report samples papers from the 2016 and 2017 GCCCE conferences that are related to learning analytics, looking for clues of current practice. The study concludes that in spite of government policy incentives to promote Internet + and big data in education, Chinese educational institutions are just at the beginning of making sense how to benefit from data driven educational practices.

Introduction

The international learning analytics (LA) research community concede that the aim of learning analytics (LA) is to use educational data to improve learning and the context in which it occurs (LAK17, 2011). To improve learning or the learning context through analytics presupposes a defined pedagogical ambition of some specificity and some ideas of a set of data to be accessed. In principle, a LA systems will use data from any available source that can serve as a proxy for a pedagogical concept under study. We then have a simple conceptual framework consisting of three entities we could use to explore the current state-of-affairs in a particular educational setting. What are the pedagogical aims? What are the data used? And what are the proxies, the intermediary concepts that bridge between the data and pedagogical objectives?

It is our assumption that in settings where LA are well established one will more easily identify the sources of data used; and what proxies are used to build knowledge about specific pedagogical objectives. For example, if the aim is to boost engagement in learning, a LA solution could use log-in instances to a LMS as a proxy for engagement and log-in information as the data for analysis. In a mature LA development community, we would expect to find discussions about the relevance of this type of data for assessing engagement and ideas for alternative data sets to be used.

In our example of measuring engagement, a software sensor was used to collect log-in data. There are, however, a great number of other sensors available in sensor-based platforms available for learning support, e.g., accelerometer, air pollution sensors, barometer, blood glucose meter, bluetooth, camera, compass, electro cardiogram, electrodermal activity meter, electroencephalogram, electromyography sensor, force gauge, galvanic skin response sensor, global positioning system, global system for mobile (GSM), gyroskope, humistor, infrared camera, microphone, near field communication receiver, radio frequency identification receiver, sonar, wifi, etc. These sensors can help with sensing activity, context, environment, and physiological state (Schneider et al. 2015). At the current stage of development, it seems that we have access to more sensor data than to sensible questions of what should be induced from the gathered data to improve learning and teaching.

Early in 2015 this author visited The Digital Lab of Open Learning at Shanghai Open University. One of the key technologies demonstrated was a electroencephalogram (EEG) scanner producing a nice set of brain wave diagrams with potential relevance for learning support. Two year later, at LAK17, the 7th international LA & Knowledge Conference in Vancouver. Mills et al. (2017) presented a paper titled "Put your thinking cap on" addressing the challenge of assessing students mental effort: "are they in deep thought, struggling to overcome an impasse, or are they zoned out?". The authors claim that EEG has not yet been utilized as a way to optimise instructional strategies; however, they had found that an intelligent tutoring system could use EEG to estimate a student's cognitive load which are correlated with learning performance (Milles et al., 2017).

When being shown the EEG scanner at the SOU Digital Lab for the second time in May 2017 the author felt an urgent need to try to connect the dots between educational data - proxies for educational concerns, and educational aims in order to get a picture of the state-of-arts in use of educational big data in China.

Learning analytics in China

Since 2015 the Chinese Government with its Internet Plus strategy has promoted application of cloud computing and big data. Educational big data is a concept that gives political resonance, and often more research specific terminology is subsumed under this umbrella term. In China numbers are per definition big, and there is no reason why not also the educational sector should produce big data. However, data for learning analytics have to originate from sensors and ICT systems of sort, and as far as we have experience through study visits and interaction with Chinese colleagues, Chinese LA practice till now focus a great deal on data traditionally found in Chinese university student management systems. These are historical data showing the result of teaching and learning already carried out. In order to make predictions based on these data one needs additional data sources describing the learner's engagement, motivation, learning activities, etc. We have observed that the Beijing Advanced Innovation Center for Future Education is exploring how sensor data from health and sports wearables could be used in LA. While these data might be easy to access (providing cooperation from the learners) it is not clear how information on health status, geolocation, physical movements, etc. would serve as proxies for learning activities. This challenge is shared by the whole research community. As Yun et al. (2016) observes: "relating a specific learner's behavior and state, represented via complex interleaved concepts such as emotion, cognition, motivation or meta-cognition (in addition to user actions) has not

yet been thoroughly investigated by either discipline especially not with the perspective of feeding this information back to the user in order to support his self-regulatory processes".

In a country as big and versatile as China, individual observations always come with a warning of being incomplete. In April 2016, George Siemens, one of the founders of LA as a research field, wrote a blog post reflection of the state of learning analytics. He said: "My attention is now primarily focused on two areas: developing LA as a field in China and increasing the sophistication of data collection. Recent visits to China, Tsinghua University and Beijing Normal University as well as an Intel LA event in Hangzhou in fall, have made it clear to me that LA is robust, active, and sophisticated in China. In many of the projects and products that I've seen, they're well ahead of where the current state of publishing in English suggests that we are." (Siemens, 2016).

Siemens' comment seems to be premised on the assumption that there are published research in Chinese language demonstrating a robust, active and sophisticated practice in LA. If that is the case, one would expect to find research evidence that connect the dots between data, proxies and instructional concepts in the main proceedings of the major Chinese yearly conference on computers in education.

Methodology

The proceedings of the Global Chinese Conference on Computers in Education are available for download at the website of Beijing Normal University's Advanced Innovation Center for Future Education, the organiser of the June 2017 conference (http://aic-fe.bnu.edu.cn/gccce2017/). The 2016 proceedings (conference theme: Learning and Teaching in the Big Data Era) and the 2017 proceedings (conference theme: The Education Reform in the Era of "Internet Plus") were downloaded.

GCCCE is an annual international academic conference organized by the Global Chinese Society for Computers in Education. According to the organisers, "the conference has been developed as a premier academic event for researchers, practitioners and policy makers in the Chinese communities for the worldwide dissemination and sharing of ideas for research in the field of Computers in Education". GCCCE papers are written in Chinese (with a few exceptions); however, the paper template mandates abstract and keywords also written in English. This allows a non-Chinese reader to select papers with keywords of interest, and in turn, run the selected papers through a machine translation service (e.g., Google Translate) to get an English version of the selected papers. Machine translation has improved a lot the last years; however, this type of translations does not give a perfect result. Nevertheless, we would claim that the machine translations are good enough to glean information about structure of the paper, what methodology is used, what data sources are analysed, and a host of other information that an experienced professional in the field would look for in academic writing.

The analysis of the papers follow usual qualitative literature review methods.

This study selected papers from the 2016 and 2017 GCCCE conferences, based on the assumption that this conference would be the most important outlet for research on learning analytics and educational big data in China. (The conference also draws contributions from Taiwan, Singapore, Hong Kong and other regions; and this is reflected in the analysis.)

Results

The GCCE 2016 conference explored "Learning and teaching in the Big data era". A sub-conference, C7, had the theme 學習分析及評量 Learning analysis and assessment. Among 225 full and short papers, and poster papers (34 of these papers in the C7 sub-conference) less than 20 papers were related to LA and big data. A review of these papers allows them to be classified in two main groups: A group of papers concerned about review of literature; contributing high level and rather abstract opinions on big data in general and potential benefits applied to education; and conceptual reasoning on the theme of big data and learning analytics without any empirical foundation. Another group of papers were focusing on very specific and advanced appliance of LA, i.e., use of EEG, face detection, and wearable brainwave scanning.

The analysis of the 2016 papers was done as a pre-study, to develop questions and analytical categories for the main study of the 2017 papers, which was published in July 2017. The pre-study raised a number of questions to both practice and research within the field of LA in a Chinese context. Big data was raised as an important concept and something that should be exploited also in the educational domain according to policy directives. However, the papers did not show any sign of actual use of big data which could have influence on educational practice.

The 2017 GCCE conference had chosen to build on the mainland China central policy slogan of the Internet Plus Era, exploring educational reform based on the affordances of cloud computing and big data. 200 papers in the three categories of full paper, short paper and poster paper were accepted. 24 papers were accepted in the C7 sub-conference, now titled Learning Analysis, Assessment, and Artificial Intelligence in Education.

The main proceedings are 862 pages; '学习分析' (learning analytics or analysis) appear in 40 pages, '大数据' (big data) in 83 pages, 'learning analytics' in 13 pages, 'learning analysis' in 10 pages, 'big data' in 9 pages, 'quantitative analysis' in 4 pages, 'data asset' in 2 pages, 'learning behavior analysis' in 2 pages, and 'analysis of learning' in 2 pages. All these pages were examined to find papers focusing directly or indirectly on learning analytics. All papers with 'learning analytics/analysis' and 'big data included in keywords were selected. In total, this selection process gave a list of 17 papers. These papers were copied out of the main proceedings document and submitted for machine translation using Google Translate.

The papers were analysed for topic, method used, research question and data used for analysis (either for the paper itself, or for the LA tool or practice reported in the paper).

Of the 17 papers two were by Taiwanese authors; one paper analysing available data to see locations of potential markets for new remedial institutions in North Taiwan, the other focusing on how to improve and introduce use of learning management systems (LMS). Three papers were literature reviews, two of them aiming at making sense of LA as a phenomenon, and the third exploring research on eye tracking in China. One paper looked into factors affecting the success of MOOCs, using survey and interview data related to a specific MOOC course. Five papers were mainly conceptual, engaging in theoretical

discussions on how small scale MOOCs differ from small scale MOOCS; how 'educational data asset' could be a useful concept in primary and middle school in the Big Data Era; how interactive classrooms are related to big data; how "learning analytics is the future"; how "researchers need to think deeply"; and how a application and evaluation index system could be constructed based on a data model of data on behaviour, teacher and student information, learning place, learning objects, and learning results. Only two paper could be said to address the use of educational data for analysis. The one conceptual paper mentioned above on the application and evaluation index system features a theoretical discussion on available data for analysis in an open online university. Another paper used data from a bespoke math platform to design and test a model looking at factors as speed, quantity and quality for instructional design. In five papers, it could be implied that the data source in mind was data from LMS or from e-schoolbag system. The one paper using e-schoolbag system data explored how a latent trait model could be used to go beyond traditional test results.

Given the target group of the GEEEC conference also including practitioners and policy makers, it should be noted that in our selection of papers there were no papers written from a implementation or practice perspective, nor from a policy maker perspective.

Discussion

Educational reform in the era of Internet plus means to connect the dots between learning activity data points, intermediary pedagogical concepts and educational aims defined in the curricula. This is not an abstract discussion on the importance of using educational big data, but tedious work engaging with teachers and students, doing research and making the research relevant to learning tools and learning design developers. If this work is commenced in a country or a region we should be able to trace it through paper contributions to the most important scholarly conferences.

Learning analytics is on the agenda of the GCCCE conference, both thematically and more practical in a sub-conference dedicated to LA and Assessment, and in 2017 also Artificial Intelligence in Education. Our analysis of the contributions, however, shows that there is a gap between the promises made in the themes and titles of the conference and the actual reports that are made in the papers. It is questionable if there is any extensive use of big data; and there is certainly no sign of learning analytics driven reform.

What data is used to feed analytics in the educational contexts described in the GCCCE papers we have analysed? It seems that the data mainly comes from LMS systems (e.g., Moodle), and it is the traditional activity log data in addition to the online test data that are positioned as candidates for analysis. This leaves out a lot of other educational data sources available to an educational institution, e.g., student information systems data, library data, data from use of other online resources than the LMS, social media data, laboratory instrument use data, etc.

In contrast to the rather traditional focus on LMS in the GCCCE proceedings analysed, there is a strain of research interest into very advanced technologies like eye tracking. This interest seems to be mainly technology driven. As the literature review by Jiang et al. (2017) of

Chinese research in 2012 - 2016 concludes, current research is mainly based on gathering experimental data demonstrating evidence of activities. However, there is no deep analysis of how this relates to effective learning, Jiang et al. claims.

Conclusions

To a Western observer everything in China is big, so big that one never knows if what one sees is just a small bit of the puzzle, impossible to generalise from. When Big Data is become a global transformative meme used by politicians to create reform one certainly expects China to be the place where the numbers are played out, – as in Educational Big Data. The reality may be different. It may well be that Siemens (2016) has found pockets of robust, active, and sophisticated LA practice in China, but this is not the full picture – at least not if we should judge from our analysis of GCCCE proceeding for the two last years. LA and Educational Big data is on the agenda for Chinese researchers; however, there is a long way to go to connect the dots between the data gathered and the analysis necessary to make learning and teaching more effective. There is a lot of data. And there are advanced sensors that create even more data. But there is a need for an active and robust research into how these data connect to the educational objectives set by the central and local authorities, and maintained on a daily basis by millions and millions of Chinese teachers.

This study is based on literature review of papers written in Chinese. The hurdles that need to be overcome in order to make sense of these data are discussed in the methodology section of this paper. Given the interest outside of China to learn and know what goes on in an area like learning analytics, it is justified to try to find ways to cross the language barrier to learn more about the current state-of-affairs and contribute to cross-cultural exchange of knowledge and experience. Chinese educational research outnumbers any other country when it comes to how many researchers, students and projects are working to reform the way we use data in education. Therefore, further work on the questions raised in this paper should be carried out analysing research contributions published in the full-text database for Chinese academic writing (cnki.com.cn).

Literature

LAK11 (2011). Call for Papers. 1st International Conference on Learning Analytics and Knowledge. Online at https://tekri.athabascau.ca/analytics/call-papers. Accessed: 2017-07-18

Mills, C., Fridman, I., Soussou, W., Waghray, D., Olney, A. M., & D'Mello, S. K. (2017). Put your thinking cap on (pp. 80–89). Presented at the the Seventh International Learning Analytics & Knowledge Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/3027385.3027431

Siemens, G. (2016). Reflecting on Learning Analytics and SoLAR. Blog post of 28 April 2016. Online at: http://www.elearnspace.org/blog/2016/04/28/reflecting-on-learning-analytics-and-solar/. Accessed 2017-07-18.

Schneider, J., Börner, D., van Rosmalen, P., & Specht, M. (2015). Augmenting the Senses: A Review on Sensor-Based Learning Support. Sensors, 15(2), 4097–4133. http://doi.org/10.3390/s150204097

Yun, H., Domanska, M., Fortenbacher, A., Ghomi, M., & Pinkwart, N. (2016). Sensor Data for Learning Support: Achievements, Open Questions and Opportunities. Proceedings of DeLFI Workshops, 1–12.

江波*，王小霞，刘迎春，高明 (2017). 我国教育领域眼动研究进展:2012-2016 (Eye Tracking in Education Research in China: A Survey from 2012-2016). In Chang, M., Jong, M., Chan, TW., Yu, S., Wu, F., Li, B., Chen, W., Hsu, CK.TC., Chuang, TY., Hsu, CY ., Zhou, Y., Liu, M., Chen, G., Chang, CC. , & Tu, S. (Eds.). (2017). Conference Proceedings of the 21st Global Chinese Conference on *Computers in Education 2017*. *Beijing: Beijing Normal* University.