Large Scale Predictive Process Mining and Analytics of University Degree Course Data

Jurgen Schulte[1], Pedro Fernandez de Mendonca[1],
Roberto Martinez-Maldonado[2], Simon Buckingham Shum[2]
University of Technology Sydney - Faculty of Science (SciMERIT)[1], Connected Intelligence Centre[2]
P.O. Box 123, Ultimo 2007, Australia
(Jurgen.Schulte, Pedro. FernandezdeMendonca, Roberto.Martinez-Maldonado, Simon.BuckinghamShum)@uts.edu

ABSTRACT
For students, in particular freshmen, the degree pathway from semester to semester is not that transparent, although students have a reasonable idea what courses are expected to be taken each semester. An often-pondered question by students is: "what can I expect in the next semester?" More precisely, given the commitment and engagement I presented in this particular course and the respective performance I achieved, can I expect a similar outcome in the next semester in the particular course I selected? Are the demands and expectations in this course much higher so that I need to adjust my commitment and engagement and overall workload if I expect a similar outcome? Is it better to drop a course to manage expectations rather than to (predictably) fail, and perhaps have to leave the degree altogether? Degree and course advisors and student support units find it challenging to provide evidence based advise to students. This paper presents research into educational process mining and student data analytics in a whole university scale approach with the aim of providing insight into the degree pathway questions raised above. The beta-version of our course level degree pathway tool has been used to shed light for university staff and students alike into our university's 1,300 degrees and associated 6 million course enrolments over the past 20 years.

CCS Concepts
• Applied computing --- Enterprise computing --- Enterprise modelling
• Computing methodologies --- Modelling and simulation --- Simulation types and techniques --- Visual analytics

Keywords
Process mining; learning analytics; predictive modeling; educational data mining, educational process visualization.

1. INTRODUCTION
Learning is a pathway, comprising the main actors (students, academics), events (courses) and outcomes (marks), which in a degree occur in a designed sequence, within a limited timeline. As cohorts of students pass through courses to complete a degree, their performance varies, depending on their academic ability and how well they study. However, the swiftness with which they progress also depends on how well individual courses are taught, how well courses are aligned horizontally as well as vertically, and how well sequences of courses are placed within the degree course structure. Naturally then, the degree structure has two views and two related goals; students' desire to pass through a degree in an optimized way (maximum academic performance, minimum time) and universities' commitment to deliver the best mix of disciplinary knowledge within degree time limits while maintaining acceptable retention rates.

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

Degree performance is monitored by both university business units and degree coordinators. From a business unit perspective the degree input-output performance (level and rate of intake, and graduation) and the retention rate are key degree performance indicators. The degree coordinator has the faculty or school interest in view, that is, teaching of disciplinary knowledge ought to have ideal scaffolding so that a student at an initially accepted entry level has a fair opportunity to pass through the degree, following a recommended degree pathway, at an appropriate pace. Ideally, the degree coordinator has intimate knowledge of the material being taught in each course, and how disciplinary knowledge and skills learned in one semester are further developed in the following semester. Often though, this kind of course level overview is difficult to achieve.

Attempts to mine educational processes of student cohorts so far have been limited to small cohorts, such a single small degree course of a few hundred students or sub-major of such degree [1]. There has been mixed success in extracting meaningful insights through educational process mining [2], one of the major obstacles being the volume of student data, even when limiting datasets to sub-cohorts of degree courses. Other challenges encountered have been profound heterogeneity and complexity within datasets and concept drifting [3]. All this together make it difficult interpret the outcome and diminishes the value of intelligence that can be drawn from it

Our goal was to uncover statistically significant and meaningful patterns in students' course pathway choices, and to provide student support units, degree and course coordinators with longitudinal indicators that could be used to inform student advice. The envisioned indicators could also help to support the streamlining of course and subject content. The hypothesis is that the more effectively students can be informed about what effort it could take them (individually) to master future subjects and stages in their course, the better their study experience will be, with better student retention rates.

2. METHODOLOGY
Student enrollment performance and progression data were provided by the UTS Warehouse and Business Intelligence division following required security and ethics clearance (UTS HREC REF NO. ETH16-0338).

The dataset allowed us to mine historical data of 1,300 UTS degree courses and some 16,000 course units taken by over 300,000 students in over 6 million enrolments. To our knowledge, this makes it the largest educational progression pathway study undertaken so far. We employed data and process mining approaches [3], using tools such as KNIME, Disco and PRoM (see [2,3] for details), data modelling in R, and RStudio's Shiny server for visualization and end user access (Figure 1) to demonstrate a proof of concept of a virtually real-time comprehensible presentation of complex processes in large scale educational data.

3. DISCUSSION

It became apparent that the data would present challenges to the envisaged process mining techniques. Firstly, students may repeat courses, choose electives offered by different degrees, change degrees or take time off to come back some later time (Figure 2). The process mining with PRoM and analysis with Disco lead to rather convoluted processes and an inflated number of processes resulting in a difficult to interpret 'spaghetti' diagram.

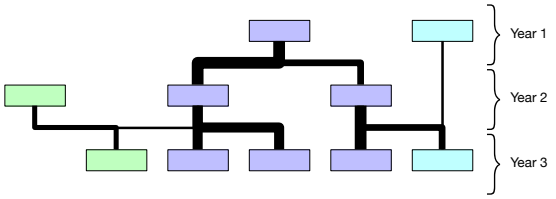

Figure 2: Simplified representation of degree complexity. Boxes represent different courses in a degree, and colors the faculties/schools by whom they are offered. Line thickness represents a relative measure of student flow density.

Secondly, with a 6 million enrolments in the dataset, it was inevitable for it to contain irregularities, as well as display historical drift as degree and course names changed (Figure 3). Education relies on innovation and re-inventing itself as society evolves and expectations change.

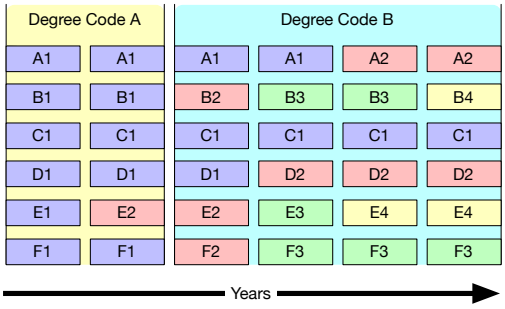

Figure 3: Example of degree and concept drift. Of the six courses (A to F) in Degree A one changes its name in year 2 (E1 → E2), Degree A then changes its name in year 3 to Degree B while largely maintaining its overall structure, followed by further course evolutions.

Process mining is particularly sensitive to such 'concept drift', but in education, this is common, and indeed, often desirable for courses to evolve. We found that these drift and process complexity issues caused an almost exponential escalation in the computing time required to extract meaningful sequences. Hence, it is perhaps not surprising that educational process mining has to date confined itself to small data samples and proof of principle studies.

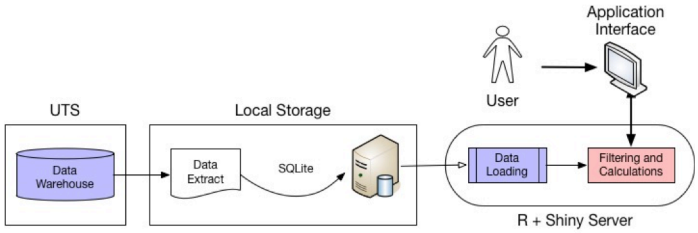

Figure 1: Data and process mining system layout.

Consequently, we employed a heterogeneity and drift resilient data mining and analytics approach. We employed KNIME to interrogate the data for categorizable structure, drift sources and patterns, and data wrangling for the production of a conveniently accessible database for the statistical analytics tool (R). This permitted process visualisation directly from raw data and scaled well. Even with the whole university dataset, this approach permitted virtually real-time, interactive interrogation and visualization, of both vertical and lateral degree course pathways by a range of customers, e.g., students, student support units, degree planners, course coordinators (Figure 4).

4. CONCLUSION

While originally promising, 20 years' data on course pathways proved intractable for conventional process mining, but was amenable to pre-structured statistical analytics, with implementation of an interactive querying and visualization tool. The project is now preparing to extend the historical cohort data representation to multistep (upward semester) forward projections and forecasting. This will serve then as a base for the development of meaningful degree health and course choice indicators.

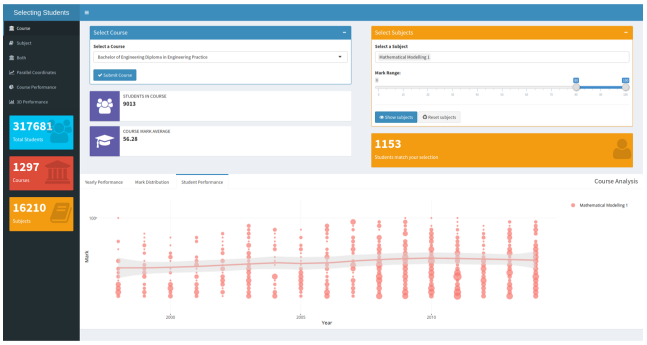

Figure 4: Snapshot of cohort course pathway predictor window of the real-time mining tool.

A following pilot with academics and university support units will assist us to assess the tool's potential to provide new insights into progression choices, levels of success, and meaningfulness of pathway indicators. Ultimately, our hope is to provide direct guidance to students.

5. REFERENCES

[1] C. Romero, S. Ventura, "Educational data science in massive open online courses," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2016.

[2] Peña-Ayala, A. 2014, 'Educational data mining: A survey and a data mining-based analysis of recent works', *Expert Systems with Applications*, vol. 41, no. 4 PART 1, pp. 1432–1462

[3] A. H. Cairns, B. Gueni, M. Fhima, A. Cairns, S. David, N. Khelifa, "Process Mining in the Education Domain," International Journal on Advances in Intelligent Systems, vol. 8 (1&2), pp. 219-233, 2015.