Eindrapportage

CurriM: Curriculum Mining

Pvenoerende instelling:
Technische Universiteit Eindhoven

Partner instellingen:

Datum:
4 october 2012

Auteur:
Mykola Pechenizkiy
Inhoud
1. Korte Samenvatting ................................................................. 3
2. Doelstelling en Aanpak ................................................................. 3
3. Kostenoverzicht ........................................................................... 4
4. Resultaten .................................................................................. 5
5. Conclusies en Geleerde lessen ..................................................... 7
6. Continuering .............................................................................. 8
7. Overige opmerkingen ................................................................. 8
1. **Korte Samenvatting**

CurriM project was aimed at investigating the potential, challenges and feasibility of the educational process mining, and in particular – of curriculum mining. By curriculum we mean both the range of courses from which students choose what courses to study, and a specific learning program that puts certain types of constrains on how students are expected to take the courses. By mining we mean a variety of both the discovery and conformance checking tasks. One of the important practical aspects included finding intuitive and convincing ways of visualizing patterns observed in the historical data, projecting them to (the user friendly abstraction of) the (formally) modeled curriculum.

Following the suggestions of the learning analytics program advisory board we focused on addressing the needs of the individual students at the first place. To achieve the stated goal we developed a CurriM tool concept illustrating how students can plan and monitor the progress of their studies, and in particular get a better understanding how the courses relate to each other, how good their personal study plan is and why it is advisable to take or focus on certain courses.

We also conducted a case study on the anonymized students’ performance data collected over past several years at TU Eindhoven. Our experimental evaluation of the existing data mining and process mining tools for the purpose of the academic curriculum modeling helped us to identify promising direction for further development of the tool and main challenges with the data quality to be addressed in the future.

2. **Doelstelling en Aanpak**

**Objectives.** The main objective of CurriM project is to help individual students (as well as educators), to get a holistic view on the study programs and to better understand the envisioned and monitor the actual educational process.

Particular kinds of questions, for which we want to provide an automated support, include but are not limited to:

- "What are the recommended choices in the curriculum for me?"
- "Why should I put attention to a particular course or prerequisites? How does a course relate to the program? Also with respect to its prerequisites and follow up dependencies?"
- "Is my study plan compliant with the official curriculum (constraints)?"
- "Am I on track, will I graduate successfully? in time? above average? with honors?"
- "Should I take now courses A, B, C or D?"

The goal of this project is to make the first practical steps towards having user friendly, intuitive and interactive educational software tools, built on (but hidden from the end users) solid foundations in process mining and data mining technology.

Another important aspect is that we want to provide evidence analyzing that from the past, which will give an insight on a particular dependency or recommendation. E.g. a prerequisite may suggest to pass the Algorithms I course with grade 8 which will give a much higher chance (90% vs. 50% success rate for those who got less than 8) for passing Advanced algorithms from the first attempt and thus do not delay related studies. We believe (and feedback from students confirm this) that such simple argumentation will help them to rethink how to distribute their studying efforts.

**Background.** Curriculum mining includes three main kinds of tasks: (i) constructing complete and compact academic curriculum models that are able to reproduce the observed behavior of students (curriculum model construction or discovery), (ii) checking whether the modeled and expected behavior matches the planned and observed behavior of students (curriculum model conformance checking), and (iii) projecting information extracted from the logs onto the model, to make the tacit knowledge explicit and facilitate better understanding of the particular academic processes (curriculum model extension) as illustrated in Figure 1.
**Approach:** In this pilot project we focused on the following tasks. Task 1: CurriM software prototype development (as a ProM plugin, see [http://www.processmining.org/prom/](http://www.processmining.org/prom/)). Task 2: Case study including pattern mining and recommendations as two main subparts. Existing techniques have been tried out on anonymized students’ performance data collected at TUE over past 12 years. Task 3: Relating tasks 1 and 2 at the levels of formalisms, data representation, and interfaces; identifying and working out intuitive examples to illustrate the concept. Task 4: Reporting

**Project period:** 01 March 2012 – 30 September 2012

**Project team:**
*Project leader:*  
dr. Mykola Pechenizkiy

*Technology experts:*  
Prof. dr. Paul De Bra (Human-computer interaction and databases), dr. Toon Calders (Pattern mining), dr. Nikola Trcka (EPM framework), dr. Boudewijn van Dongen (process mining), dr. Eric Verbeek (ProM software)

*PhD students:*  
Pedro Toledo (Software prototype as ProM plugin), Aekyung Kim (Recommenders)

*Domain experts:*  
STU, University level: Dr. Karen Ali (Hoofd Onderwijs en Studenten Service Centrum) and Drs. Marcel Visschers (Students performance database)  
Department of Computer Science: Dr. Marloes van Lierop and Prof. Dr. Mark de Berg, Directors of the bachelor and graduate program.

### 3. Kostenoverzicht

The costs include mainly the hours we put on this project and minor expenses like traveling to Surf meetings, which are included into the project management category. Because of the refocusing of the project (based on the feedback of the advisory board on the submitted proposal) to target individual students as the main user group we introduced changes (as little as possible) into the initial planning. More hours we put on the development of the software tool concept and the experimental study was extended to cover the course recommendation. To compensate for this, the amount of reporting activities has been reduced.

The summary of the costs is enclosed in the table below.  
TU/e account number: 15.82.49.658 (Rabobank). Please use the following reference for the money transfer: 10014743 (Surf 15.892).
4. Resultaten

The main achieved results correspond to the CurriM software prototype development (Task 1) and the experimental evaluation of the direct applicability of existing data mining and process mining tools for academic curriculum modeling at the Department of Computer Science, TU Eindhoven for which real data was used (Task 2).

CurriM has been built as a ProM plugin to allow for a direct application of many process mining tools and for embedding all the standard tools for data preprocessing and unification, process model extraction and conformance checking available in ProM and applicable to the educational data. This also allows data/process mining experts to study the applicability of advanced techniques to CurriM modeling tasks.

Figure 2 gives a general idea how the current CurriM GUI looks like. A short demo of the GUI has been recorded and made available at http://www.win.tue.nl/~mpechen/projects/edm/CurriMDemo.mp4. The software report will be made available as an online technical report.

Note from Figure 2 that all the technical details are hidden from the user. E.g. curriculum model is represented as a set of codified constraints with Colored Petri net and LTL. A simple choice of 2 courses out 3 by a student that can be visualized as a selection of corresponding cells at the back end is represented as a pattern shown in Figure 3. This is the representation that can be used by the conformance checker of ProM for automated check whether a student obeys this curriculum constraint or not.
Figure 2 – One of the CurriM screenshots presenting the student’s current curriculum choice of the student (all with grey background), currently examined course (selection with blue borderline), mandatory prerequisites (red borderline) one of which is satisfied and the other one is not, and its relation to one of the follow up courses (green borderline).

Figure 3 – Example 2-out-of-3 Pattern Check using Colored Petri net formalism

The case study was focused on pattern mining and recommendations as two main tasks. Curriculum model discovery from the log and model checking became our secondary priority as these concern directly mainly educators responsible, and students mainly in an indirect way. Existing techniques have been tried out on anonymized students’ performance data collected at TUE over past 12 years.

The experimental results on TUE dataset suggest that the recommendations for the courses to be taken and predictions of time to completion of the studies are highly indicative and not always accurate. This can be attributed as to the quality of the analyzed data as to the possible necessity to further develop corresponding predictive modeling techniques. This requires further research. We reflect more on the lessons learnt in the following section.

However, even with noisy and inaccurate data we worked with, we could discover useful evidence explaining the importance of the prerequisites and motivate the importance of particular courses in the curriculum. E.g. a typical type of question we were asking the pattern mining tool - what are the courses such that if students have problems with the course $A$, it will be likely that students will have problems with some other course(s) $B$. Several of such patterns have been identified for the computer science related programs.
5. Conclusies en Geleerde lessen

Conclusions: Curriculum mining is recognized to be a promising endeavor to undertake. In CurriM project we worked out the concept, illustrated its potential and feasibility, thus making an important step towards bringing educational data mining technology to the practical use in the university education.

Early feedback: Several students that were shown the concept did like it. They consider CurriM as a tool that provides (1) orientation in the curriculum, allowing to explore the connections and dependencies between different courses, (2) awareness of how good their personal education route is, where they currently are, where they are heading to, how well they do in comparison with others, why a particular course or prerequisite is important, and (3) recommendations on what is advisable to take next. Overall, this enables better planning and regular monitoring of their own progress, concentrating on subjects that look important, not just interesting.

It goes without saying that the education management is enthusiastic when they see the potential of engaging students into the education process, encouraging them to better orient themselves in the process and finding motivation and objective arguments to follow the curriculum making conscious choices along the path. However, the education management also can see clearly the potential of extending the functionality of the tool to answer the key questions they are interested in. These include, but are not limited to:

- “What is the real academic curriculum (study program)?”
- “How do students really study?”
- “Is there a typical (or the best) way to study?”
- “Do current prerequisites make sense?”
- “Are the particular curriculum constraints obeyed?”
- “How likely is it that a student will finish the studies successfully or will drop out?”

As we have had these insights already from the start of the project, this was taken into consideration of CurriM design, as we consider such extensions to be feasible also.

Lessons learnt:
Learning analytics and educational data mining heavily rely on data and its quality determines whether outcome of the curriculum modeling is meaningful or not. We use a generic approach allowing for use of a standardized input like e.g. MXML format of ProM tool. Basic data like student_id, course_id, semester_id, and grade can be easily imported from any database that stores such information. So, if we take an ideal situation where the data is complete, consistent and clean, we can hope for good results to come out.

But when it comes to practice we need to deal with lots of challenges. In case of TUE as with many other institutions, the data reflects the corresponding use practices relevant for a particular period in time. That data collection and management processes were good enough for the administrative needs, but not necessarily good for curriculum modeling. To give a few examples – at TUE the curriculum changed from semesters to trimesters to semesters to quartiles in the past several years. So, course codes change over the years. These changes are not mapped and codified. Therefore, it is difficult to keep all the data of one course together. As the mappings of course ids from older to newer curriculum have not been put into a knowledge base, it is also not possible to make an automated reasoning about such changes. On the other hand, one can ask herself whether courses of 1992 can be compared to courses of 2012. It is likely that content of courses changes substantially over time and it is given in different ways by different teachers. A domain expert should be able to map the changes of the course codes each year, or remove courses for which can be said that they are not relevant for the later years.

Another typical difficulty is that for quite some students, it is not known whether they finished their study successful or dropped or suspended their studies. Until the very recent past in the database only the (first) attempt for an exam would be the first signal that the student actually attempted to follow the corresponding course. These issues are not important for a transactional administrative database, but become a serious barrier for the automated data analysis tools.
Besides, the data without meta-data or knowledge about the educational processes, present in completely separate sources like study guides (explicit), and in heads of the responsible educators (implicit) can easily mislead any intelligent data analysis tool. There are study guides, but no formal curriculum model can be easily constructed by each faculty for each program. It is essential to have right tools developed to make the curriculum modeling easy and transparent for the education management and the faculty in general.

6. **Continuing**

The project leader Mykola Pechenizkiy ([http://www.win.tue.nl/~mpechen/](http://www.win.tue.nl/~mpechen/)) can be contacted for any further information regarding the achieved results in this project and about the envisioned possibilities for the continuation.

The problem of curriculum modeling has been studied scarcely in a few areas with not much evidence that this can be done effectively and conveniently for the expected users. Most of the papers that can be found in the literature focus on particular tasks like student performance or student drop out prediction, recommendation of learning materials for the courses, finding frequent paths through the curriculum, discovering actual curriculum of the study program and performing conformance checking with process mining techniques and alike. But those papers that appear do gain some popularity. E.g. our modest case study on drop out prediction at EE department of TUE in 2009 has been cited more than 30 times in 2 years.

The CurriM project idea was presented and discussed with peers at EDM 2012 ([http://educationaldatamining.org/EDM2012/uploads/procs/Posters/edm2012_poster_11.pdf](http://educationaldatamining.org/EDM2012/uploads/procs/Posters/edm2012_poster_11.pdf)) and it was well received. Now we plan to summarize the results obtained so far and submit them for presentation or demonstration at LAK 2013 and/or EDM 2013. We plan also to prepare an extended journal version describing the sketch of the CurriM developed software platform along with the use cases illustrating the connection between the intuitive user interface with the formal representation of the curriculum at the back end, the use of process mining, data mining and recommender systems techniques for providing awareness and recommendations for individual students and academics responsible for the organization, monitoring and enhancement of the educational process.

Overall, we consider this pilot to be a success. An important step has been made towards bringing educational data mining technology to the use in the university education. However, it should be understood that there are no quick wins in CurriM in the sense that a few month project won’t provide the required technology, infrastructure and support for adopting curriculum mining. There is a substantial amount of work to be done to bring curriculum mining to daily practice of a complex organization such as university. This includes but not limited to: working out the full cycle of the information flow consisting of the pattern mining, predictions and recommendations, and curriculum model enrichment pipelines; working out different views and functionality for students vs. educators, HCI/usability aspects; improving data quality; facilitate knowledge base construction (meta-data, mappings); facilitating curriculum formalization for the faculties (tooling). Last, but not least, CurriM integration with the existing institutional infrastructure would be a serious project on itself. Therefore, at the moment we consider it as a development of a stand-alone tool to be promoted for use in parallel with the existing infrastructure and with an access to the existing university data repository.

7. **Overige opmerkingen**

I find the organization and especially the personal support of John Doves in this programme to be great. However, the overall organizational overhead would fit nicely to a 12 months 1.0 fte or larger project. For the size of the project we had had, the overhead was excessive (even though it was all relevant and useful).