A WORKFLOW MANAGEMENT TOOL TO SUPPORT THE DEVELOPMENT OF OPEN LEARNING ANALYTICS APPLICATIONS

Fabiano Ammirata, Giovanni Fulantelli, Concetta La Mattina, Davide Taibi

National Research Council of Italy, Institute for Educational Technology (ITALY)

Abstract

Most of the approaches of Learning Analytics rely on a predefined set of indicators, and most of the tools used to analyse learning traces focus on a specific aspect, thus lacking in flexibility to provide multiple analysis. In this perspective, Learning Analytics tools must be more flexible in order to support personalized Learning Analytics approaches in which indicators are defined by users. In this paper, we present a workflow management tool aimed at supporting the development of Learning Analytics applications specifically designed to deal with students' interactions in online learning management systems. A use case in which the tool has been used to detect students at risk of dropout is also illustrated.

Keywords: learning analytics, educational data analysis, workflow management system.

1 INTRODUCTION

The huge amount of traces left by students when using online learning platforms has paved the way to the development of new research fields. Educational Data Mining (EDM) first, and Learning Analytics (LA) few years later, represent two major research areas that use data to obtain new insights on learning processes.

Definitions of EDM and LA suggest that they share the same general goal to support education "by improving the assessment, how problems in education are understood and how intervention are planned and selected" [1]. At a lower logical level, data analysis methods in EDM are increasingly influencing Learning Analytics techniques. Even though differences between EDM and LA exist, as reported by Siemens and Baker [1], to the aim of this paper we focus on Learning Analytics, but keeping the Educational Data Mining as a source of data analysis techniques and models to feed Learning Analytics approaches. Most of the approaches of LA rely on a predefined set of indicators [2][3], and most of the tools used to analyse learning traces are focused on a specific aspect thus lacking in flexibility to provide multiple analysis. In this perspective, LA tools must be more flexible in order to support personalized LA approaches in which indicators are defined by users, thus enabling self-reflections in the definition of goals and research questions to be addressed.

In this paper we present a workflow management tool aimed at supporting the development of Learning Analytics applications specifically designed to deal with students' interactions in online learning management systems.

The objectives of this platform are:

- Interlinking data made available from different sources in order to facilitate the development of applications and services learner oriented;
- Providing a flexible tool to define personalized Learning Analytics approaches;
- Supporting the creation of Learning Analytics models, according to user self-defined goals and research questions.
- Supporting actors with different roles that can work collaboratively in the design of the elaboration workflow according to their expertise.

Moreover, the environment presented in this paper takes up the challenge of providing a set of tools that require low effort for non-technical users but at the same time enabling easier access to advanced functionality for expert users [1].

The paper is structured as follow: next section introduces a brief literature review on methods for data analysis; then, in section 3 we introduce the proposed framework and in section 4 we illustrate the

platform that has been developed in order to implement the framework; an exemplary use case is reported in section 5; lessons learnt and suggestions for future research work conclude the paper.

2 METHODS FOR DATA ANALYSIS

The methods introduced in this section provide an overview about the very large number of methods used in EDM and LA applications.

According to Baker and Inventado [4], analysis methods for EDM can be classified into four main categories: prediction models, structured discovery, relationship mining and discovery with models.

The basic idea behind the prediction models is to make conjecture about future events starting from the knowledge of other events potentially related to them. The main analysis tools belonging to this category are regression and classification. Regression is used when the variable to be predicted is defined in a continuous range of values, while classification is adopted when response variable is binary or categorical. Examples of regression are illustrated in [5], which apply linear regression to study the connections between students' language features, extracted through natural language processing (NLP), and measures of Math Identity, composed of math self-concept, interest, and value.

A typical classification problem when analysing educational data is to assess the student's knowledge at a specific time in order to activate mechanisms that can improve it and inform teachers about students learning progress [8].

Examples of studies that use classification methods can be found in [6]. Feng and Heffernan in [6] formulate a Rasch model that relates results from the Massachusetts Comprehensive Assessment System (MCAS) to students' accuracy, speed, number of attempts and help-seeking behaviour. Romero et al in [7] present the results of an experimental research work carried out to compare different classification algorithms used until 2012, adopted to predict students' final marks depending on their active participation in online courses.

Another method used in Educational Data Mining is the Relationship Mining, which consists in identifying the most significant variables in a large dataset depending on the focus of the analysis and the relationships the researcher is interested.

The Structure Discovery algorithms are used to find the intrinsic structure into a set of data based on observed variables, without any ground truth or a priori idea of what should be found [4]. Clustering, factor analysis, Domain Structure Discovery are common examples of the most used Structure Discovery algorithms in EDM. Baker and Yacef [9] defined the use of an existing EDM or analytics model as a component in a new EDM or analytics analysis "discovery with models". The Discovery with Models is less common with respect to the others, but its use has grown during time in EDM application. Hershkovitz et al. in [10] illustrate discovery with models method with a case study.

The platform presented in this paper is based on a framework that is not bounded to a specific set of methods, but creates an abstract layer in which different methods can be implemented. The following sections provide a detailed description of the framework.

3 THE FRAMEWORK

The framework at the basis of the educational workflow management platform presented in this paper aims at answering the four questions posed by Chatti et al. in their Learning Analytics reference model [11]: What kind of data does the system gather, manage, and use for the analysis? Who is targeted by the analysis? Why does the system analyse the collected data? How does the system perform the analysis of the collected data?

At a higher level, the framework presents a flow-based, modular and web-based environment that allows the users to choose their own customized tools for data analysis.

Specifically, the framework is structured in three phases: extraction, processing and visualization. The first phase deals with all the preliminary operations such as data extraction from specific sources and at a scheduled period of time; the second one deals with manipulating data through different elaboration blocks; the third and last phase deals with the data visualization.

The potential beneficiaries of LA solutions belong to different categories of users, whose interests in the results of LA can vary significantly. We distinguish three categories of users with different roles:

- administrators, who define the LA modules to be enabled, assign roles to other users, establish access and editing privileges.
- data managers, who define the process based upon the modules enabled by the administrator
- data users, who access and visualize the produced data.

3.1 Data Extraction

As stated by Siemens et al. in [12], analytics need to be multi-sourced. In fact, students' traces are left in different contexts: during their online activities (e.g. in searching for information [13], commenting in social media, participating in e-learning courses); but also in traditional settings (e.g. during a city tour through the use of GPS), in educational activities requiring physical abilities (e.g. gym exercises through the use of accelerometer), and so on. These traces provide educators with useful insights on the learning processes and shed lights on how these processes can be improved.

In this perspective, Learning Analytics engines have to implement specific procedures to import data from both online activities and physical world-data, thus leveraging methods from the LA and EDM research fields.

Because of the multiplicity of sources, data must be refined before being processed. For this reason, extraction modules in a LA engine have to standardize data according to a specific data model. After that, standardized data coming from the extraction modules can be passed through the processing modules that follow in the elaboration pipeline.

3.2 Data Processing

The core element of a LA engine is the data processing stage. Data received from the data extraction module are elaborated according to one or more statistical methods, whose choice mainly depends on the objectives of the LA strategy and on the available data. Elaborated data are then passed to the data visualization module.

As reported in section 2, enumerating all the possible techniques to model and analyse educational data is almost impossible; accordingly, the efficiency of a Learning Analytics engine should not be assessed against the number of data processing models implemented in it, rather on the possibility to add new components that implement specific data processing models. Flexibility and scalability are therefore two important requirements for this stage of the framework, and the possibility to add new data processing models on demand is one of the main objectives.

3.3 Data Visualization

After the data processing phase, the last step in the proposed framework is data visualization. Visualization is a critical point of all the learning analytics, because it is crucial to transform a set of data into visual information that can be easily interpreted by a user. Learning dashboards are used to visualize students' traces during their learning activities. They are not only used to monitor learning activities that take place in online contexts, but they are also useful to provide an overview of the class activities also in face-to-face and blended educational settings. As stated by Dillenbourg et al. in [14] dashboards represent tools that provide evidences and support teachers' analysis and decisions, but they do not take decisions. Furthermore, dashboards support students in visualizing their learning progress and improve awareness and self-reflection.

Dashboards provide users (both teachers and students) with visual representations of specific indicators related to the performance of learners. Examples of indicators are scores in exercises, quizzes, or other forms of assessments, as well as time spent to accomplish a specific learning task. A variety of charts has been used as visualization hints. The type of charts to be used is strictly dependent on the type of indicator to be shown.

4 THE EDUCATIONAL DATA WORKFLOW ENVIRONMENT

The framework defined in the previous sections has guided the development of a software environment aimed at supporting the three fundamental phases of data extraction, processing and visualization, by taking into account the specific requirements that underpin the conception of the framework: easy-of-use, replicability, collaboration. Specifically, the environment presented in this paper has been designed as a workflow management system that provides users with an infrastructure to visually design data

analysis and elaboration pipeline. The visual approach, backed by an easy-to-use interface, supports the re-use of existing pipelines, allowing users with different roles and different expertise to work collaboratively in the data elaboration process.

The workflow management system consists of two main sub-systems. The first is focused on the data extraction and processing, while the second is specialized in the data visualization. The Educational Data Workflow environment is based upon the Open Source project Node-Red. Node-Red provides a browser-based editor that supports the definition of elaboration pipelines. Node-Red has been originally designed for the Internet of Things (IoT). Starting from the original project, the core elements consisting of a basic set of modules have been reused, and new modules specifically designed to easily extract, process and visualize educational data have been developed. Moreover, Node-Red provides a multi-user environment, which facilitate the implementation of collaboration processes between users with different roles and performing different set of operations.

One of the features of Node-RED is its flow-based architecture. All functionalities are therefore individually implemented in separated modules, each of which manages an input payload to return a processed output.

The runtime of Node-red is based on Node.js to fully exploit its event-driven and non-blocking model. In addition, Node.js is well suited to the dual use of the client-server service because JavaScript is used for both the backend and the frontend. Furthermore, its native asynchronous behaviour and the possibility of easily adding new modules, make Node-Red a power environment to implement all the phases of data extraction, processing and visualization.

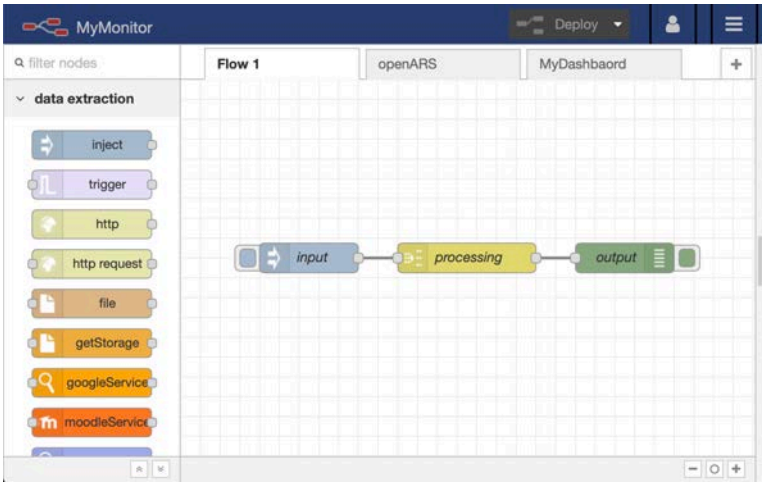

Figure 1. The workflow management design panel.

Figure 1 shows the main panel of the workflow environment, as it is presented to users after the login. On the left column, the modules that can be used in the execution pipeline are listed. An interesting approach carried out by the Node-Red platform regards the elaboration process design: the visual interface allows users to simply drag the modules from the left columns into the central part of the panel and connect them in order to create the elaboration pipeline. (Fig. 1 shows three modules performing a simple operation). Each of the modules in the pipeline implements a task which is triggered upon arrival of new input data or when an event occurs. A search bar is also provided to support users in finding specific modules.

The deployment of the designed pipelines is accomplished through a button (located at the top right of the window, together with buttons for managing the account and to view other operations and settings).

Data that have been extracted and loaded into the framework are processed through a series of specific modules. They allow users the possibility to choose between different processing settings. Multiple simultaneous data are also supported by an asynchronous modular structure that starts multiple processes simultaneously.

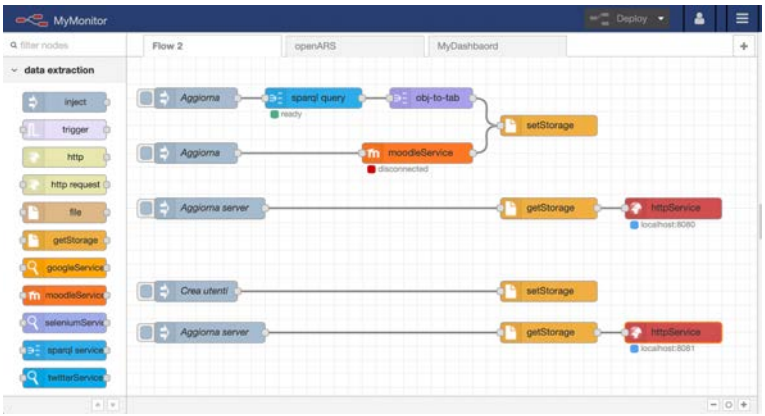

Figure 2: An elaboration pipeline.

Figure 2 represents the skeleton of a web server. Specifically, it shows pipelines of a simple web app with authentication. Input buttons will timely start processing pipelines. The output of the various processes can then be saved locally or provided via API rest using the appropriate modules. In this case, data are extracted through queries in a relational database.

The results of the pipelines are finally interpreted and visualized through a web app that connects to the server where the extraction and processing of data have occurred, collect the results, and pass them to a dashboard application. In order to make the analysis of the final data promptly available and easily accessible, a cross-platform dashboard application has been developed; it has been conceived in such a way to render data both in computers as well as in portable devices (e.g. tablets, smartphones, etc.).

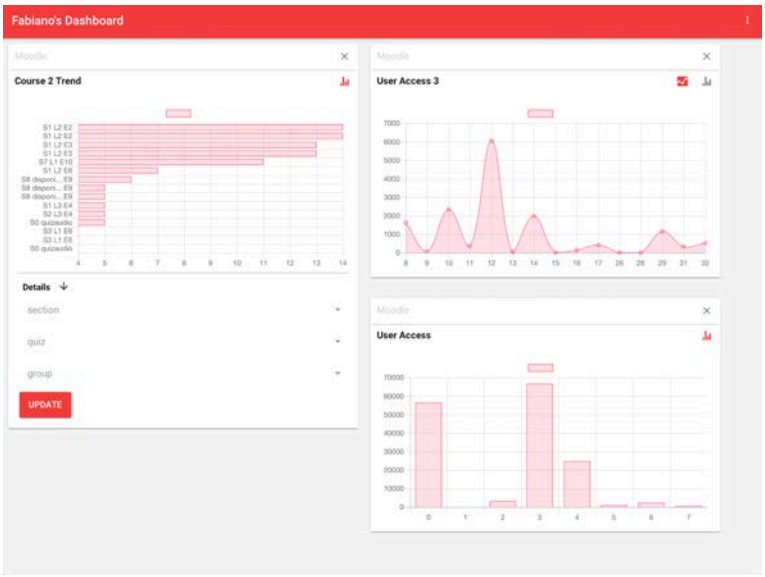

Figure 3: Data visualization in dashboards.

Figure 3 shows a dashboard consisting of three different panels, in which three different dimensions of the data are visualized. Additional panels can be easily inserted through the button at the top left of the screen. The user can choose the chart to be added from a list of available charts. This list is directly defined in the design environment.

5 USE CASE

One of the areas in which LA techniques provide relevant support is the analysis of educational data to identify students at risk of dropout.

Rosè et al. in [15] studied the evolution of the social interactions between students from the University of Pittsburgh's into Massive Open Online Courses (MOOC). In this study, a Mixed-Membership Stochastic Block model was used to follow student movements in sub-communities in order to analyse

students' dropout rates. Their analysis shows the important link between students' interaction and dropout. In particular, starting from a graphical representation of the trend over time in discussing different topics, through a survival model they observed that for students who were active in the first week of the online course was less likely to dropout the programme than for their colleagues who started their active participation later. Eckles et al. in [16] collected archival data from a student information system and, through logistic regression, proved that there is a relationship between students' retention and some social network variables. The analysis made by Robinson et al. in [17] is focused on the prediction in a MOOC system of students drop out starting from their opinions about what they will learn in this class that can be applied in their life. In order to accomplish this task, they used a logistic regression classifier and they found that demographic characteristic are important predictors of students drop out.

According to the literature presented above, we have used a synthetic grade to reassure or alarm the student when they were detected at risk of dropout. This grade is the result of a series of elaboration on the operations carried out by student during their learning activity on line. In particular, data have been analysed through a decision tree predictive model. This algorithm of machine learning uses students' observed characteristics (e.g. past academic grades, school attendance and family conditions) to classify them into three categories: pass, fail and at risk. The result of this process is shown in Figure 4, where a given colour of traffic light is assigned to each student. In particular, the traffic light is red when the student's failure is predicted, yellow when s/he is at risk and green if s/he is expected to pass.

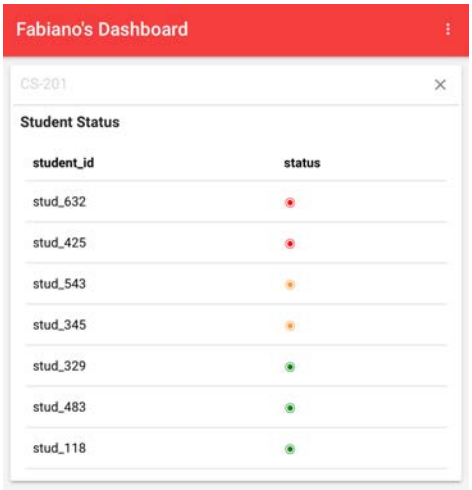

Figure 4: Data visualization.

6 CONCLUSIONS

The final aim of the project presented in this paper is the development of services and applications that provide students, teachers, educators and other educational stakeholders with an ecosystem including the tools necessary for statistical analysis of open data coming from different sources, and finalized to educational purposes. The system presented in the paper can support Learning Analytics applications along three phases: starting from the extraction of data, it performs statistical elaboration of them, even in complex execution pipelines, and provides tools for the graphic visualization of the results.

According to open science principles, the presented system can be used to define transparent, shareable, and reproducible pipelines. The modules developed are just an example of what can be developed using the presented framework. Finally, it should be noted that, because of the flexibility in introducing new modules, it is straightforward to apply the proposed framework to different fields: data journalism, education, and so on.

REFERENCES

[1] Siemens G, Baker RSJd. Learning analytics and educational data mining: towards communication and collaboration. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. Vancouver, British Columbia, Canada; 2012, 1–3.

[2] Muslim, A., Mohamed Amine Chatti, Tanmaya Mahapatra, and Ulrik Schroeder. 2016. A rule-based indicator definition tool for personalized learning analytics. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16). ACM, New York, NY, USA, 264-273.

[3] Muslim, A., Chatti, M., Mughal, M. and Schroeder, U. 2017. The Goal - Question - Indicator Approach for Personalized Learning Analytics. In Proceedings of the 9th International Conference on Computer Supported Education (CSEDU 2017) - Volume 1, pages 371-378.

[4] Baker R.S., Inventado P.S. (2014) Educational Data Mining and Learning Analytics. In: Larusson J., White B. (eds) Learning Analytics. Springer, New York, NY.

[5] Crossley, S., Ocumpaugh, J., Labrum, M., Bradfield, F., Dascalu, M., & Baker, R. S. (2018). Modeling Math Identity and Math Success through Sentiment Analysis and Linguistic Features. *International Educational Data Mining Society*.

[6] Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, *19*(3), 243-266.

[7] Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008, June). Data mining algorithms to classify students. In *Educational data mining 2008*.

[8] Pavlik Jr, P. I., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis--A New Alternative to Knowledge Tracing. *Online Submission*.

[9] Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3-17.

[10] Hershkovitz, A., de Baker, R. S. J., Gobert, J., Wixon, M., & Pedro, M. S. (2013). Discovery with models: A case study on carelessness in computer-based science inquiry. *American Behavioral Scientist*, *57*(10), 1480-1499.

[11] Chatti, M., Dyckhoff, A. L., Schroeder, U., and Thüs, H., 2012. A reference model for learning analytics. *Int. J. Technol. Enhanc. Learn. 4, 5/6 (January 2012), 318-331.*

[12] Siemens, G., Gašević, D., Haythornthwaite, C., Dawson, S., Buckingham Shum, S., Ferguson, R., Baker, R. S. (2011). *Open Learning Analytics: an integrated modularized platform*. Edmonton, AB, Canada.

[13] Taibi, D., Rogers, R., Marenzi, I., Nejdl, W., Ahmad, Q. A. I., & Fulantelli, G. (2016, May). Search as research practices on the web: the SaR-Web platform for cross-language engine results analysis. In *Proceedings of the 8th ACM Conference on Web Science* (pp. 367-369). ACM.

[14] Dillenbourg, P., Zufferey, G., Alavi, H., Jermann, P., Do-Lenh, S., Bonnard, Q., & Kaplan, F. (2011). Classroom orchestration: The third circle of usability. In CSCL2011 proceedings (Vol. 1, pp. 510-517). Hong Kong: International Society of the Learning Sciences.

[15] Rosé, C. P., Goldman, P., Zoltners Sherer, J. and Resnick, L. (2015). Supportive technologies for group discussion in MOOCs. Emerging eLearning: Vol. 2: Iss. 1, Article 5. Available at: https://scholarworks.umb.edu/ciee/vol2/iss1/5

[16] Eckles J., Stradley E. 2011. A social network analysis of student retention using archival data. Social Psychology of Education, 15 (2) (2011), pp. 165-180.

[17] Robinson C., Yeomans M., Reich J., Hulleman C., and Gehlbach H. (2016). Forecasting student achievement in MOOCs with natural language processing. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, LAK '16, pages 383–387, New York, NY, USA. ACM, 2016. doi: 10.1145/2883851. 2883932.