Towards Custom-Designed Professional Training Contents and Curriculums through Educational Process Mining

Process Mining in the Education Domain

Awatef Hicheur Cairns[1], Billel Gueni[1], Mehdi Fhima[1], Andrew Cairns[1] and Stéphane David[1]
Nasser Khelifa[2]

[1] ALTRAN Research, [2] ALTRAN Institute
Vélizy-Villacoublay, France
e-mails: {awatef.hicheurcairns, billel.gueni, mehdi.fhima, andrew.cairns, stephan.david, nasser.khelifa}@altran.com

***Abstract*—Educational Process Mining constitutes a new opportunity to better understand students' learning habits and finely analyze the complete set of educational processes. In this paper, we investigate further the potential and challenges of Process Mining in the field of professional training. Firstly, we focus on the mining and the analysis of social networks between course units or training providers. Secondly, we propose a two-step clustering approach for partitioning educational processes following key performance indicators.**

Keywords- Process Mining; Educational Data Mining; Curriculum Mining; Key Performance Indicators; ProM.

I. INTRODUCTION

Recently, education and training centers have started introducing more agility into their teaching curriculum in order to meet the fast-changing needs of the job market and meet the time-to-skill requirements. In fact, modern curriculums are no longer monolithic processes. Students can pick the courses from different specialties, may choose the order, the skills they want to develop, the level (from beginner to specialist), the way they want to learn (theoretical or practical aspects) and the time they want to spend. This need for personalized curriculum has increased with the emergence of collaborative tools and on-line training which often supplement and sometimes replace traditional face-to-face courses. The broad number of courses available and the flexibility allowed in curriculum paths could create, as a side effect, confusion and misguidance. Students may be overwhelmed by the offer and blurred on the time required to enter and remain in the job market. Moreover, teachers and educators may lose control of the education process, its end-results and feed-back. The use of information and communication technologies in the educational domain generates large amount of data, which may contain insightful information about students' profiles, the processes they went through and their examination grades. The deriving data can be explored and exploited by the stakeholders (teachers, instructors, etc.) to understand students' learning habits, the factors influencing their performance and the skills they acquired [6] [15]. Rather than relying on periodic performance tests and satisfaction surveys, exploiting historical educational data with appropriate mining techniques enables in-depth analysis of students' behaviors and motivations [6] [8]. *Educational Data Mining* (EDM) is a discipline aimed at developing specific methods to explore educational. EDM methods can be classified into two categories – (1) Statistics and visualization (e.g., Distillation of data for human judgment*)*, and (2) Web mining (e.g., Clustering, Classification, Outliers detection, Association rule mining, Sequential pattern mining and Text mining) [15]. However, most of the traditional data mining techniques focus on data or simple sequential structures rather than on full-fledged process models with concurrency patterns [20] [21]. Precisely, the basic idea of *process mining* [1] is to discover, monitor and improve real processes (i.e., not assumed nor truncated processes) by extracting knowledge from event logs recorded automatically by Information Systems. Our research aims to develop generic methods which could be applied to general education issues and more specific ones concerning professional training or e-learning fields for:

- The extraction of process-related knowledge from large education event logs, such as: process models and social networks following key performance indicators or a set of curriculum pattern templates.
- The analysis of educational processes and their conformance with established curriculum constraints, educators' hypothesis and prerequisites.
- The enhancement of educational process models with performance indicators: execution time, bottlenecks, decision point, etc.
- The personalization of educational processes via the recommendation of the best course units or learning paths to students (depending on their profiles, their preferences or their target skills) and the on-line detection of prerequisites' violations.

In this paper, we focus mainly on (1) process model discovery, deriving from Key Performance Indicators; (2) social network discovery between training courses and training providers. We used the ProM framework (i.e., a "pluggable" environment for process mining) [7] for process discovery and analysis from event logs. For the first time, to our knowledge, the present approach handles a professional training dataset of a consulting company involved in the training of professionals. The rest of this paper is organized

 ISBN: 978-1-61208-364-3

as follows. Section II introduces process mining techniques and their application in the educational domain. Section III presents our approach for social networks mining and process models discovery. Section IV describes briefly the PHIDIAS platform. Finally, section V concludes the paper.

II. EDUCATIONAL PROCESS MINING

The purpose of Process Mining is to develop automated techniques to extract process-related knowledge from *event logs* [1]. An event log corresponds to a set of process instances following a business process. Each recorded event refers to an *activity* and is related to a particular process instance. An event can have a *timestamp* and a *performer* (i.e. a person or a device executing or initiating the activity). Moreover, in such logs, events are assumed to be totally ordered. The scope of Process Mining includes process discovery (i.e., extracting process models from an event log), conformance checking (i.e., monitoring deviations by comparing model and log), social network/organizational mining, automated construction of simulation models, model extension, model repair, case prediction, and history-based recommendations. Educational Process Mining (EPM) or Curriculum Mining refers to the application of process mining techniques in the education domain [20] [21]. Beyond limitations of EDM, EPM enables greater insights into underlying educational processes. To illustrate, process mining techniques were used by Pechenizkiy et al. [13] to investigate the students' behavior during online multiple choice examinations. Southavilay et al. [18] used process model discovery techniques to mine and analyze a collaborative writing process. Analysis techniques were also applied to check the conformance of a set of predefined constraints (e.g., prerequisites) with event logs [21]. However, the application of Process Mining techniques in the education domain faces numerous challenges related to event logs' specificities:

Voluminous Data: event logs in the education domain, particularly those coming from e-learning environments, contain massive amounts of fine granular events and process-related data. Real-life testing showed that most of the current process mining techniques/tools are unable to handle massive event logs [12] [14] [17].

Heterogeneity and complexity: educational processes are complex and flexible by nature reflecting the high diversity of behaviors in students' learning paths. Consequently, traditional process discovery techniques generate intricate models (spaghetti) which are often very confusing and difficult to analyze [22].

Concept drift: in the education domain, subjacent curriculum and trainings may evolve over time and occasionally undergo major changes. Concept drift refers to a situation where the process will change while being analyzed [5].

Interpretation of results by the end users: visualization techniques and notation simplification is a major stake to facilitate interpretation by the end users, using suitable academic notation or lists of recommendations [14].

III. CASE STUDY: ANALIZING TRAINING PROCESSES USING EPM TECHNIQUES

A. Data description and preprocessing phase

In our case study, the dataset encompasses the *employees' profiles*, their *careers* (i.e., their jobs/assignments history) and their *training paths* (performance and satisfaction surveys throughout the different training phases). The data being scattered in several databases, we had first to rebuild a consolidated event log (using an ETL -*Extract, Transform and Load*) containing the following information: *Employee Id, Training Id, Timestamp, Training provider Id, Training Cost* and the *List* of all the *employees' missions* over a three years period. Secondly, we transformed this event log into MXML (Mining eXtensible Markup Language) format by using the *ProM Import* plug-in [7], with the condition that (1) an employee identifier corresponds to a process instance identifier (i.e., an employee training path corresponds to a process instance), (2) a training identifier corresponds to a task identifier tagging 'start' and 'end' events, with various attributes (grades, satisfaction, employee profile, etc.).

B. Social network mining

Social Network Analysis (SNA) refers to the collection of methods, techniques and tools in sociometry aiming at the analysis of the structure and composition of ties in social networks [4]. In our case, we conducted mining and analysis of the key interaction patterns between training providers and training courses, using social mining techniques deriving from the process mining field. These techniques aim to extract social networks from event logs based on the observed interactions between performers and depending on how process instances are routed between these performers [1] [3]. These interactions can be generated following one of these five metrics: (1) *handover of work*, (2) delegation or *subcontracting* of tasks, (3) frequent collaboration (*working together*) (4) similarity in executed tasks and (5) *reassignment* of tasks. In order to generate social networks between training courses, we replaced originator IDs by training IDs of the same events (i.e., trainings) during the event log conversion step in *ProM import*. Social mining plug-ins generate graphs where each node represents a training provider (resp. a training course) where the names have been anonymized for privacy reasons. The oval shape of the nodes in a graph (see Figures 1 and 3) visually expresses the intensity of in and out connections (arrows) between the nodes: a higher proportion of ingoing (outgoing) arcs lead to greater vertical (horizontal) distortion of the oval shape (see Figures 2 and 4). We use different views (a ranking view, a stretch by degree ratio,

 ISBN: 978-1-61208-364-3

etc.) and two key SNA indicators when generating these graphs, depending on the patterns we want to extract. The key SNA indicators [4] we used are: (1) *Degree Centrality* of a node (i.e. the number of nodes that are connected to it): it represents the popularity of a node (e.g., training courses or training providers) in a community (e.g., training paths or curriculums). (2) *Betweenness Centrality* of a node: representative of the enabling power of a node to connect two different groups (i.e., two different training paths or curriculums). A node (i.e., training provider or training course) with high betweenness centrality value means that it performs a crucial role in the network. We apply these five metrics to mine social networks between training providers and training courses, giving the following outcome:

1) Handover of work: within a process instance, there is a handover of work from individual *i* to individual *j* if there are two subsequent activities where the first is completed by *i* and the second by *j*. In our case, this metric allowed us to discover the flow of trainees (specified by the direction of the arrows) between training providers and courses. For instance, in Figure 1, two providers are connected if one performs a training causally followed by a training performed by the other provider. We distinguished two groups of providers strongly related to each other (clustered in cliques) following their causal involvement in training paths. Training providers without arc are those which offer very stand alone trainings without causal dependency with others. In Figure 2, the most important training courses (trainings with ID 4 and ID 1) appeared to play a central roles in training paths. In Figure 3, the size of training courses (with high betweenness) indicates their crucial role as a bridge (i.e., intermediate trainings) between different types of trainings. We can deduce that:

- Training providers or courses with *high degree* are the most popular ones. Training providers or courses with no connection with others represent outliers, providing very specific skills, not involved in training paths.

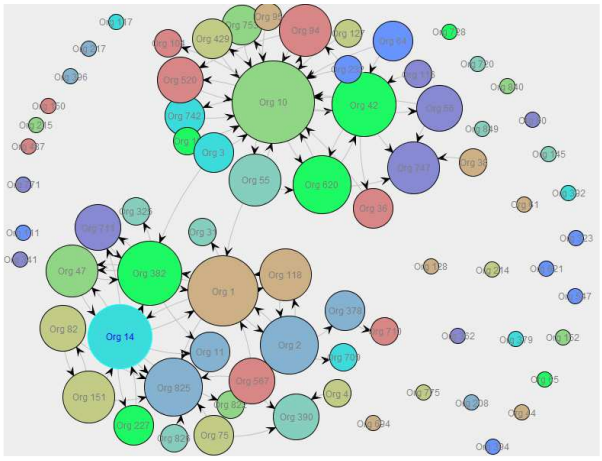

Figure 1. Social network showing handovers between providers of the top 80% of followed training courses using a *size by ranking* view.

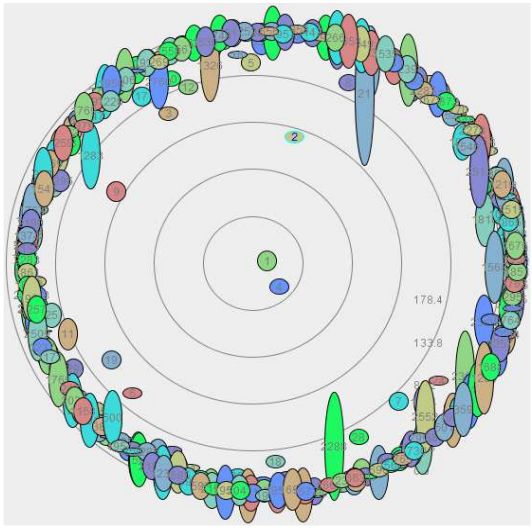

Figure 2. Social network showing *handovers* between training courses using (1) *a rancking view on degree* and (2) a *stretch by degree ratio*.

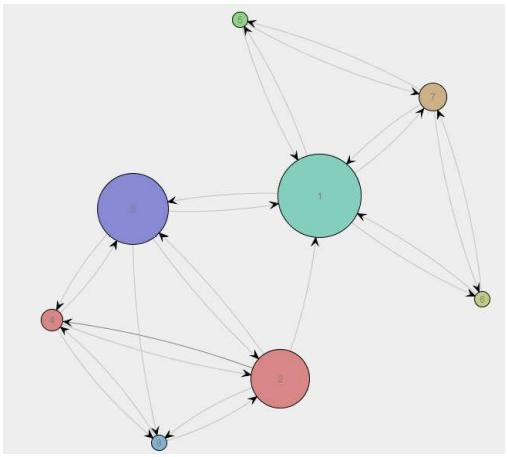

Figure 3. Social network showing *handovers* between the top 60% of followed training courses, using a *ranking on betweenness centrality* and a *size by ranking* view.

- Nodes with no incoming arcs are training providers (or training courses) who only initiate learning processes (i.e., give the basics), while nodes with no outgoing arcs are training providers (or courses) who perform only final trainings (i.e., complete training paths with the most required skills).
- Training courses strongly connected to each other hint popular or typical curriculums (or learning paths). The direction of the edges gives the order of training courses followed by students in such curriculums.
- Training courses or providers with *high betweenness centrality* represent the ones playing a crucial role as a bridge (i.e., offering intermediate trainings) connecting different types of learning paths.

 ISBN: 978-1-61208-364-3

2) Subcontracting metric: A resource *i* subcontracts a resource *j*, when in-between two activities executed by *i* there is an activity executed by *j*. In this case, the start node of an arc represents a contractor and the end node means a subcontractor (see Figures 4 and 5). In this case study, this metric allow us to extract complementary patterns between training courses and providers. Using SNA measures, we deduce that:

- Nodes (i.e., training providers or courses) with a high out-degree of centrality (indicated by a horizontal oval shapes) usually play the role of contractors (the main providers or trainings which give basic skills in these training paths).
- Nodes (i.e., training providers or courses) with a high in-degree of centrality (indicated by a vertical oval shapes) usually act as subcontractors (providers or trainings which give complementary notions or skills allowing to enhance the notions given by contractors in these training paths).

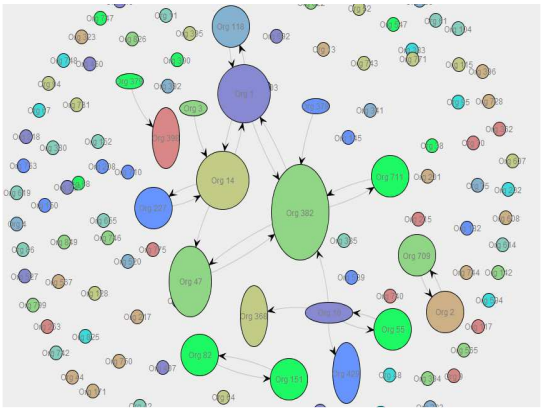

Figure 4. Social network showing *subcontracting* between training providers of the top 90% of followed training courses.

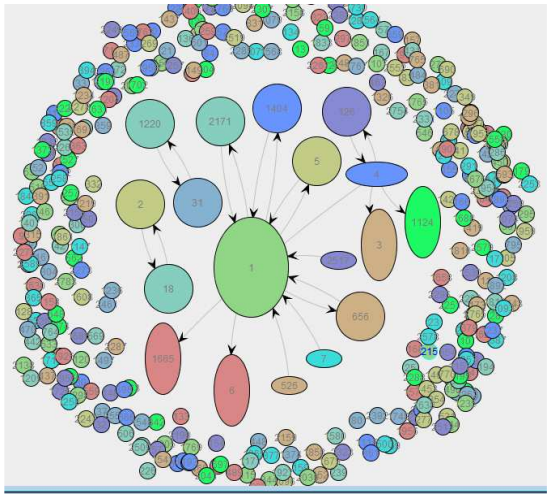

Figure 5. Social network showing *subcontractin*g between the top 80% of followed training courses.

3) Working together metric: This metric ignores causal dependencies but simply counts how frequently two recourses are performing activities for the same case (see Figure 6). We can deduce from this social network the most popular curriculums (training providers or courses that work together i.e., are involved together in training paths). The difference with the handover metric is that the latter gives us the order followed by students in such curriculums.

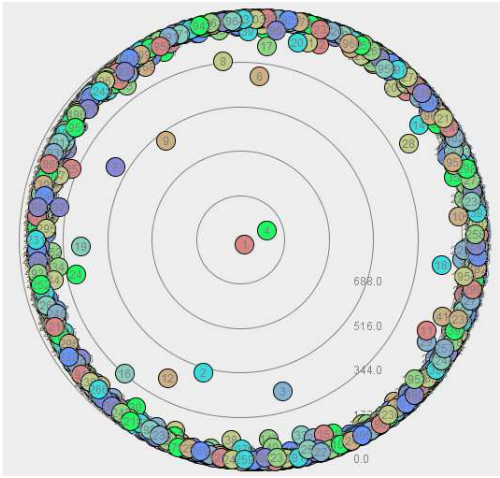

Figure 6. Social network based on *working together* between training courses using *a rancking view on degree*.

4) Similar task metric: This metric determines who performs the same type of activities in different cases. In our case study, this metric makes sense only to generate relationship between training providers. Therefore, it allows us to detect training providers who perform the same kind of trainings in curriculums.

This experience shows that social network analysis based on event logs is a powerful tool for analyzing coordination patterns between training courses and training providers. Such an approach can also be used to mine interesting patterns about students' behaviors in on-line environments based on resources' usage logs and various interaction logs (e.g., in the case of an intelligent tutoring system).

C. Process discovery using Clustering Techniques

Clustering techniques can be used as a preprocessing step to handle large and heterogeneous event logs by dividing an event log into homogenous subsets of cases following their similarity. One can then discover simpler process models for each cluster. For this purpose, several clustering techniques have been developed and implemented in ProM [11], such as the Trace Clustering plug-in [9] [17], the Sequence Clustering plug-in [22] and other clustering approaches based on time [11]. Clustering of event logs still remains a subjective technique. A desired goal would be to introduce some objectivity in partitioning the log into homogenous cases. In this paper, in order to identify the best training paths, we propose a two-step clustering approach where

 ISBN: 978-1-61208-364-3

training paths are firstly partitioned following a performance indicator then they are partitioned following their structural similarity. *The first step* consists of creating clusters of similar trainees' profiles based on a training path performance indicator expressed via two criteria: (1) *employability* (matching degree between skills required by a mission and skills obtained via training) and (2) *duration* between the training end and new job start. Based on trainees' profiles, three clusters are created using the k-means technique. The optimal number of these clusters (three) is determined using a method based on the average silhouette of many clustering where the number of the clusters is varied [16] [19]. For more details on this method we refer the reader to [10]. Figure 7(a)-(b) presents, respectively, the clusters we obtained and the silhouette used to determine the optimal number of clusters.

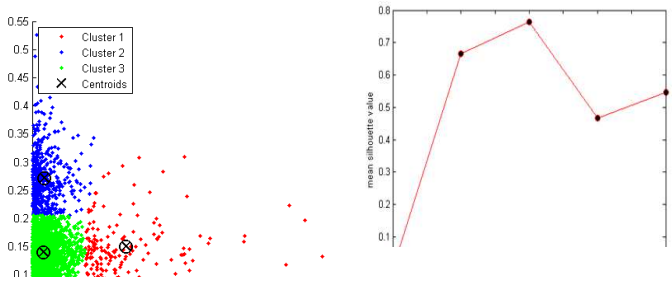

Figure 7. (a) Three clusters obtained (b) silhouette used to determines the optimal number of clusters

Let us note that the first cluster groups trainees with the best employability factor and the shortest duration between trainings' end and new missions. Cluster 2 and Cluster 3 group less optimal training paths regarding employability factor and period of unemployment. We use the fuzzy miner plug-in of ProM (given its robustness to noises) to discover the process model from the training traces of the trainees grouped in the first cluster. We obtain clearly identifiable training paths, as illustrated in Figure 8. Let us note that these training paths correspond to the highest performing ones regarding employability factor and period of unemployment.

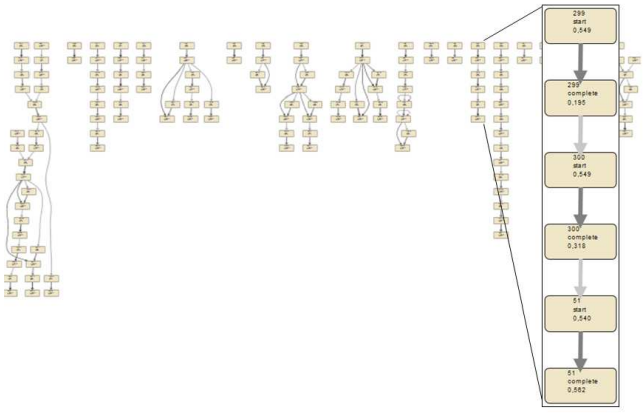

Figure 8. A fragment of the process model showing all the training patterns of cluster 1

In the second step, for further simplification, we group training paths (traces in log events) from each of the clusters discovered in the first step, following their structural similarity using the Sequence clustering technique proposed by Veiga and Ferreira [22]. Each cluster is based on a probabilistic model, namely a first-order Markov chain. The sequence clustering technique is known to generate simpler models than trace clustering techniques developed in [22]. In our example, when we apply the sequence clustering technique on the second group of trainees with an average employability (i.e., the second cluster of the first step), we obtain three more clusters (cluster 2.1, cluster 2.2 and cluster 2.3). Figure 9 shows the training model obtained from the cluster 2.1 obtained above, where only transitions occurring above the threshold of 0.05 are represented.

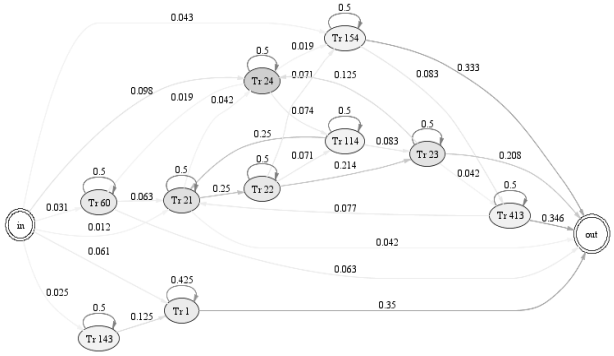

Figure 9. The process model describing the training paths of the first cluster of the second group of trainees (Cluster 2.1), with an edge threshold of 0.05

IV. PHIDIAS: A PLATFORM FOR DISTRIBUTED EDUCATIONAL PROCESS MINING

To implement our approach, we aim to develop an interactive platform tailored for educational process discovery and analysis. This platform will allow different education centers and institutions to load their data and access advanced data mining and process mining services. Such a platform has to address several issues related to: (1) the heterogeneity of the applications and the data sources; (2) the connection to some web portals and desktop applications to allow users dealing with the data and exploiting analysis results; (3) the ability to add new data sources and analysis services; (4) the possibility to distribute heavy analysis computations on many processing nodes in order to optimize and enhance platform response time. To reach these targets we adopt a Service Oriented Architecture using an Enterprise Services Bus (ESB) depicted in Figure 10. This architecture is composed of the following elements: data sources, Enterprise Service Bus, business applications and tools, web services, web portals and connectors. The core of this architecture is the application bus which guarantees the interoperability and integration of the data sources and applications. We have chosen to use ESB architecture in order to have a flexible architecture allowing easily plugging of new applications, data sources and web portals.

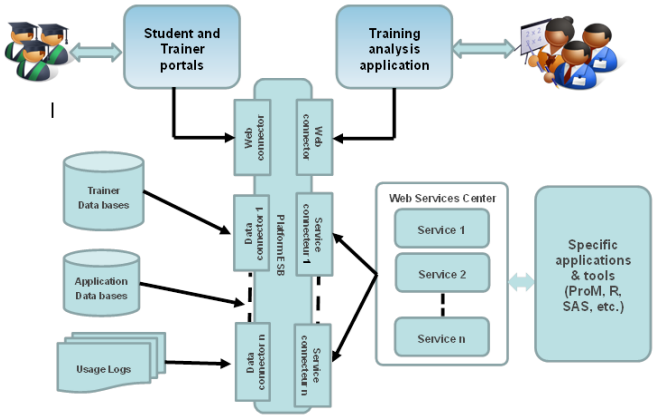

Figure 10. PHIDIAS Architecture

V. CONCLUSION AND FUTUR WORK

In this paper, we showed how social mining techniques can be used to examine interactions between training providers and training courses, involved in students' training paths. We also proposed a two-step clustering approach to extract the best training paths depending on an employability indicator. Our future work will continue in several directions. Firstly, we intend to combine the approach proposed to mine interaction patterns with other mining techniques which allow to discover interaction patterns between students in their collaborative learning tasks, communication actions and online discussions [2]. Secondly, we intend to apply the conformance checking techniques to check if prerequisites, other kinds of constraints and training path templates were indeed always respected. Thirdly, we plan to investigate further clustering techniques in event logs partitioning to extract typical or atypical training paths depending on domain specific performance indicators and/or on a set of predefined training path templates. Finally, the proposed architecture will be implemented and deployed and tested on a *distributed environment* connected to several data sources and applications. We plan also to conduct a case study that would illustrate the feasibility of process mining approaches in an on-line education setting.

REFERENCES

[1] W. M. P. van der Aalst et al. "Process mining manifesto," In BPM 2011 Workshops Proceedings (BPM 2011), Aug. 2011, pp. 169–194, doi:10.1007/978-3-642-28108-2_19.

[2] W. M. P. van der Aalst and Adriy Nikolov, "EMailAnalyzer: An E-Mail Mining Plug-in for the ProM Framework," BPM Center Report BPM-07-16, BPMCenter.org, 2007.

[3] W. M. P. van der Aalst and M. Song, "Mining social networks: Uncovering interaction patterns in business processes," The second International Conference on Business Process Management (BPM 2004), LNCS, vol. 3080, June 2004, pp. 244–260, doi. 10.1007/978-3-540-25970-1_16.

[4] C. Aggarwal, "An Introduction to Social Network Data Analytics," Social Network Data Analytics, Springer, 2011, pp. 1-15.

[5] R. Bose, W. M. P. van der Aalst, I. Zliobaite, and M. Pechenizkiy, "Handling Concept Drift in Process Mining," The 23rd International Conference (CAiSE 2011) LNCS 6741, Springer, June 2011, pp. 391–405, doi:10.1007/978-3-642-21640-4_30.

[6] T.Calders and M. Pechenizkiy "Introduction to The Special Section on Educational Data Mining," SIGKDD Explorations Newsletter, ACM, may 2012, pp. 3-6, doi:10.1145/2207243.2207245.

[7] B. van Dongen, H. Verbeek, A. Weijters, and W. van der Aalst, "The ProM framework: a new era in process mining tool support," The 26th International Conference (ICATPN 2005) LNCS Vol. 3536, June 2005, pp. 444–454, doi:10.1007/11494744_25.

[8] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers, "Predicting Students Drop Out: a Case Study," The 2nd International Conference on Educational Data Mining (EDM 2009), July 2009, pp. 41–50, ISBN 978-84-613-2308-1.

[9] R.P. Jagadeesh Chandra Bose and W.M.P van der Aalst, "Context Aware Trace Clustering: Towards Improving Process Mining Results," The SIAM International Conference on Data Mining (SDM 2009), April 2009, pp. 401-412.

[10] L. Kaufman and P. J. Rousseeuw. "Finding Groups in Data: An Introduction to Cluster Analysis". by Leonard Kaufman, Peter J. Rousseeuw, March 1990, ISBN: 0-471-87876-6.

[11] D. Luengo and M. Sepúlveda, "Applying Clustering in Process Mining to Find Different Versions of a Business Process That Changes over Time," The Business Process Management Workshops (BPM 2011) Aug. 2011, pp. 153-158 doi:10.1007/978-3-642-28108-2_15.

[12] J. Munoz-Gama, J. Carmona, and W.M.P. van der Aalst, "Conformance Checking in the Large: Partitioning and Topology," The 11th International Conference on Business Process Management (BPM 13), Aug. 2013, pp. 130–145, doi:10.1007/978-3-642-40176-3_11.

[13] M. Pechenizkiy, N. Trčka, E. Vasilyeva, W. van der Aalst and P. De Bra, "Process Mining Online Assessment Data," The 2nd International Conference on Educational Data Mining (EDM 2009), July 2009, pp. 279–288.

[14] M. Reichert, "Visualizing Large Business Process Models: Challenges, Techniques, Applications," The 1st Int'l Workshop on Theory and Applications of Process Visualization (BPM 2012) Sep. 2012, LNCS Vol. 132, pp. 725-736, doi:10.1007/978-3-642-36285-9_73.

[15] M., Romero and C., Ventura, "Data mining in education," The Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol 3, Feb. 2013, pp. 12–27.

[16] G. Seber, "Multivariate Observations," Hoboken, NJ: John Wiley & Sons, Inc., 1984.

[17] M. Song, C. W. Günther, and W.M.P. van der Aalst, "Trace Clustering in Process Mining," Business Process Management Workshops (BPM 2008) Sep. 2008, LNCS Vol. 17, pp 109-120, doi:10.1007/978-3-642-00328-8_11.

[18] V. Southavilay, K. Yacef, and R. A. Calvo, "Process mining to support students' collaborative writing," The 3rd International Conference on Educatinal Data Mining (EDM 2010) June 2010, pp. 257-266.

[19] H. Spath, "Cluster Dissection and Analysis: Theory," FORTRAN Programs, Examples, New York: Halsted Press, 1985.

[20] N. Trčka and M. Pechenizkiy "From Local Patterns to Global Models: Towards Domain Driven Educational Process Mining," The 9th Conference on Intelligent Systems Design and Applications (ISDA 2009), Dec. 2009, pp. 1114–1119, doi:10.1109/ISDA.2009.159.

[21] N. Trčka, M. Pechenizkiy, and W. van der Aalst, "Process Mining from Educational Data (Chapter 9)," Handbook of Educational Data Mining.. CRC Press, 2010, pp. 123–142, doi: 10.1201/b10274-11.

[22] G. M. Veiga and D. R. Ferreira,"Understanding Spaghetti Models with Sequence Clustering for ProM," Business Process Management Workshops (BPM 2009) Sep. 2009, LNCS vol. 43, pp. 92–103, doi:10.1007/978-3-642-12186-9_10.

 ISBN: 978-1-61208-364-3