Semantic-based Model Analysis towards Enhancing Information Values of Process Mining: Case Study of Learning Process Domain

Kingsley Okoye[a*], Abdel-Rahman H. Tawil[a], Usman Naeem[a], Syed Islam[a] and

Elyes Lamine[b]

[a]School of Architecture Computing & Engineering, University of East London, UK
{u0926644, A.R.Tawil, U.Naeem, Syed.Islam}@uel.ac.uk

[b]Université de Toulouse, Mines-Albi, CGI, Campus Jarlard, Albi Cedex 09, France
Elyes.Lamine@mines-albi.fr

Abstract. Process mining results can be enhanced by adding semantic knowledge to the derived models. Information discovered due to semantic enrichment of the deployed process models can be used to lift process analysis from syntactic level to a more conceptual level. The work in this paper corroborates that semantic-based process mining is a useful technique towards improving the information value of derived models from the large volume of event logs about any process domain. We use a case study of learning process to illustrate this notion. Our goal is to extract streams of event logs from a learning execution environment and describe formats that allows for mining and improved process analysis of the captured data. The approach involves mapping of the resulting learning model derived from mining event data about a learning process by semantically annotating the process elements with concepts they represent in real time using process descriptions languages, and linking them to an ontology specifically designed for representing learning processes. The semantic analysis allows the meaning of the learning objects to be enhanced through the use of property characteristics and classification of discoverable entities, to generate inference knowledge which are used to determine useful learning patterns by means of the Semantic Learning Process Mining (SLPM) algorithm - technically described as Semantic-Fuzzy Miner. To this end, we show how data from learning processes are being extracted, semantically prepared, and transformed into mining executable formats to enable prediction of individual learning patterns through further semantic analysis of the discovered models.

Keywords: process mining, semantic annotation, ontology, learning process, event logs, knowledge discovery

1. Introduction

In recent years, a common challenge in many information systems has been on how to create tools or techniques capable of providing platforms for event data exploration which derive understandable patterns as well as make the discovered patterns explicable [1]. Process mining [2] is one of the existing techniques used for pattern discovery and has been successfully applied for classical mining of processes where each process execution is recorded in terms of events log sequences. Through process mining, meaningful information about how activities depend on each other in a process domain has been made possible and has proven to be essential for extracting models capable of creating new knowledge. However, a shared challenge with most of the existing process

mining techniques is that they depend on tags in event logs information about the process, and therefore to a certain extent are limited because they lack the abstraction level required from real world perspectives. Majority of the process mining techniques in literature are purely syntactic in nature, and to this effect are somewhat ambiguous when confronted with unstructured data. This means that these techniques do not technically gain from the real knowledge (semantics) that describe the tags in event log of the domain process.

In this paper, we show that analysis provided by current process mining techniques can be improved by adding semantic information to the event log or models for any given domain process. We ascertain how the result of process mining algorithms can be enriched through semantic representation of the deployed model, inference knowledge discovery and semantic reasoning. The semantic-based analysis takes advantage of the rich semantics described in event log of a learning process domain, and links them to concepts in an ontology in order to extract useful patterns by means of semantic reasoning. Such reasoning is supported due to the formal definition of ontological concepts and expression of relationships that exist between event logs of the learning process. Our method uses the semantics of the sets of activities within the learning process to generate rules and events relating to task, to automatically discover and enhance the process model ontology through semantic annotation of the elements within the developed learning knowledge base. The semantic viewpoint is captured by annotating the elements in the systems based on two types of analyses (i) how to make use of the semantics that describes the available data? and (ii) how to mine the semantic information? [3]. The main opportunity is that the mining and process analysis outcome is enhanced as a result of it being based on concepts rather than the event tags/labels.

The rest of the paper is structured as follows; Section 2, shows how we extract the input data necessary to be mapped into a semantic model for analysis, and also provide an example of a learning process execution data which we use to demonstrate our approach throughout this paper. In section 3, we describe a learning problem use case scenario and implementation of the semantic-based approach to show the usefulness of using semantic-based approach to manage perspectives of process mining. Section 4 explains the semantic formalisation and mining algorithm. In section 5, we discussed and analysed appropriate related works in this area of research, and finally conclude the paper and point out directions for future research in section 6.

2. Mining Event Data: Case study of a Research Process Domain

The purpose for designing the Semantic Learning Process Mining algorithm is to extract, semantically prepare and transform event data of a learning process into mining executable formats that allows us to perform an improved learning process analysis, and then build a semantic model to represent the deployed model. The first step towards achieving this goal is to capture event data about a learning process (case study of research process domain) and generate process mapping to show in details how the learning activities has been performed and to reveal interesting connections between the different elements (process instances). The approach allows us to further perform an advanced analysis of the learning data. The described technique involves the extraction

of process history data from the learning execution environment, which is then followed by submitting the resulting event streams format to the process analytic environment for mining and analysis at a more conceptual level. We focus on revealing information about resources hidden within the event logs, the individual process elements that makes up the learning model, and identifying useful arrays towards enriching the information values of the derived model based on the concepts. Consequently, suitable learning patterns are determined which enables the automatic creation of the learning process mapping as shown in - Fig. 1
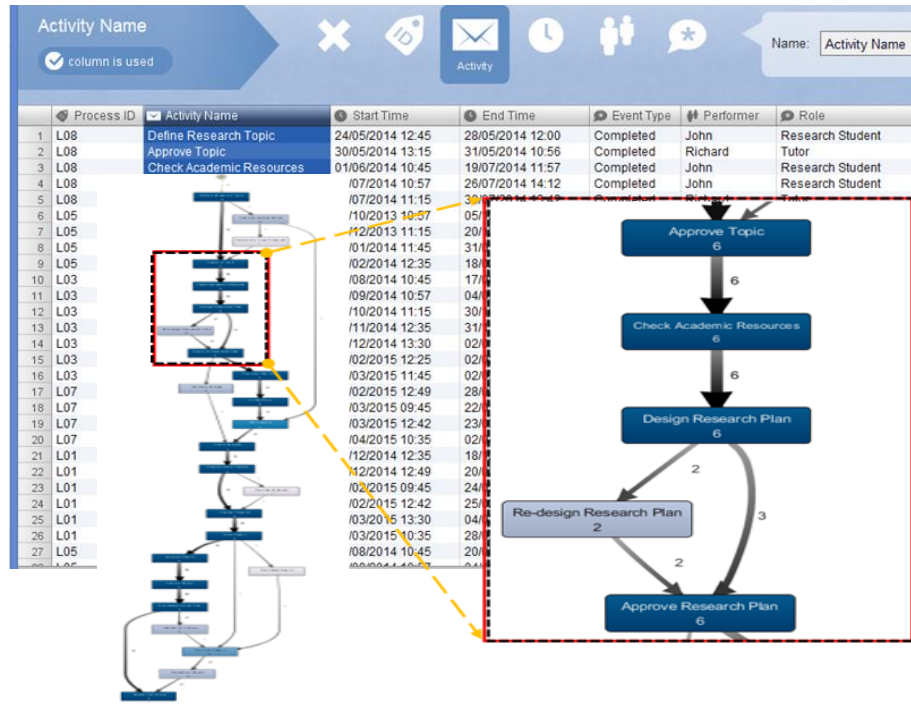
Fig. 1 Event Data and Control-flow of Mapped Process Model in Discovery miner

In Fig. 1 we show how we extract the input data necessary to be mapped unto a learning model using the *Fuzzy miner* in Disco [4]. The benefit of performing such tactics is that the resulting process map allows us to quickly and interactively explore the learning process into multiple directions and to answer concrete questions about the learning activities workflow. More importantly, the approach allows us to further model and hold inference reasoning to generate process improvement ideas along the way.

The process mapping step is necessary especially when our main purpose is to make the semantics information about the learning event data readily available for mining and analysis at conceptual level [5]. The approach reveals the process map and provides us with the opportunity to focus on the stream of learning behaviours, and to see the paths they follow in the process model. *Case id tags* are used to assign the identifier for process instances and *Activity tags* for the set of task performed during the learning

process. We associate *Timestamp tags* with activity instances for the purpose of sequencing to mark the start and end time of each event.

3. Learning Problem Use Case Example

To show the usefulness of the semantic-based approach, we define resolution towards resolving a real life question and understanding about a learning process. We use the questions LQ1 and LQ2 to provide learning knowledge representation and specification by revealing interesting connection among the learning domain process and object types, in order to provide a better understanding of how the different elements within the Learning Process Knowledge base relate and interact with each other. We show the usefulness of the tactics by answering the following question to determine;

LQ1. What attribute or paths do successful learners have in common?

LQ2. What attributes distinguishes the successful learners from the uncompleted ones?

We use running example of the research process to prove how our approach can be used to answer the learning question LQ1 and LQ2. The resulting process map discovered using the fuzzy miner algorithm in Disco (Fig. 1) shows that the research process involves the workflow of the journey from choosing the research topic to being awarded a certificate, and comprises sequence of practical steps or set of activities through which must be performed in order to find answers to the research problems. However, the deployed model still do not disclose to us how the individual process instances that makes up the model interact or differ from each other (semantic abstraction levels), which attributes they share amongst themselves within the knowledge base, or the activities they perform together or differently. For instance, who are the individuals that have successfully completed the research process? For this reason, we believe that by adding semantic knowledge to the deployed model, it will be possible for us to determine and address the above mentioned challenges as we show in the next section.

3.1 Semantic Model Mapping, Assertions and Analysis of the Learning Ontology

The approach in this paper uses semantic-based process mining and analysis to find out what paths successful researchers follow or have in common, and what attributes distinguishes the successful researchers from the uncompleted ones as developed in questions LQ1 and LQ2. The purpose is not only to answer the following question, but to show how by referring to the attributes (semantic knowledge) and the application of semantic reasoning, it becomes easy to refer to a particular case/group of learners. We develop using process descriptions in Protégé [6] (Fig. 2, 3, 4 and 5) the workflow library for the derived model (Fig. 1). We provide in the semantic model, four milestones; Defining the Topic Area → Review Literature → Addressing the Problem → Defending the Solution in order to determine and explain the steps taken during the research process. These milestones consist of sequence of activities, and the order in which these learning activities are carried out has the capability of determining the research outcome. In Fig. 2, 3, 4 and 5 we show the Learning Activity concepts that are defined in our model and how they are mapped to the various Milestones of the Research Process to ensure sequence of transitions during the learning process.
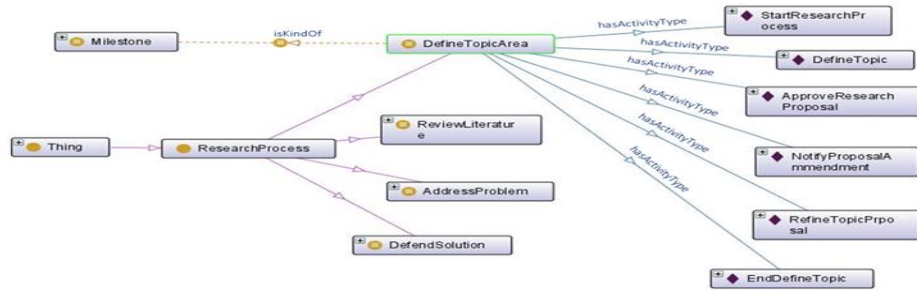
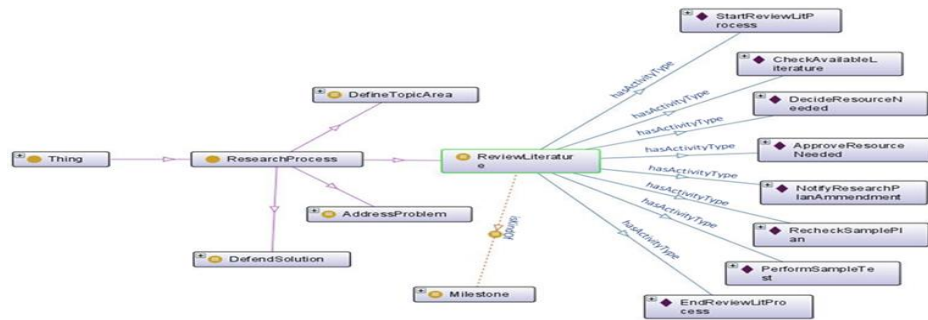Fig. 2 Ontology Graph ActivityConcept mapping for DefineTopicArea Milestone.

Fig. 3 Ontology Graph of ActivityConcept mapping for ReviewLiterature Milestone.
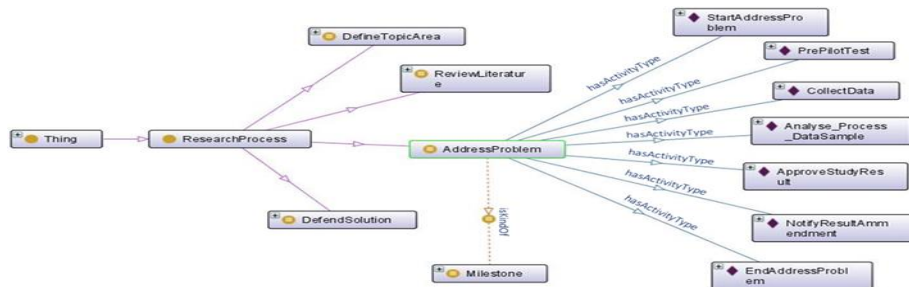

Fig. 4 Ontology Graph of ActivityConcept mapping for the AddressProblem Milestone
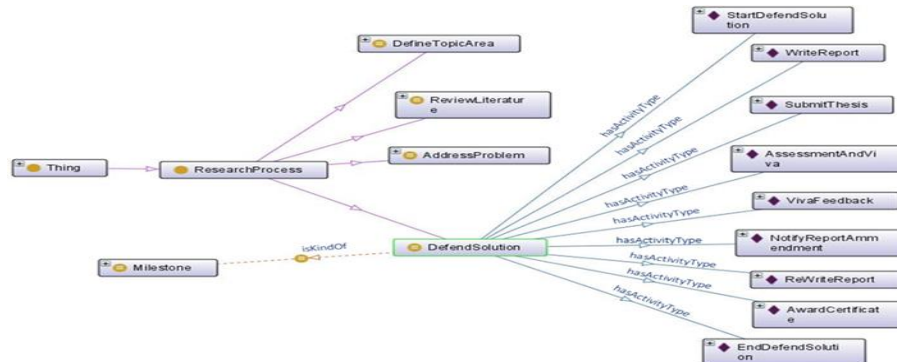

Fig. 5 Ontology Graph of ActivityConcept mapping for the DefendSolution Milestone

The purpose of semantic mapping of the learning activity concepts is that it allows the meaning of the learning objects and properties to be enhanced through the use of *property characteristics* and *classification of discoverable entities*. To address the learning problem stated in LQ1, we refer to the deployed model, and to this effect, describe that a Successful Learner is a subclass of, amongst other NamedLearnerCategory, a Person that performs some Learning Activity Concepts, who has a universal object property restriction/relationship with the four milestones of the ResearchProcess Class. As described in Fig. 6 - the *necessary condition* is: if something is a Successful Learner, it is *necessary* for it to be a participant of the Learning ActivityConcept class and *necessary* for it to have a kind of sufficiently defined condition and relationship with the four class: DefineTopicArea, ReviewLiterature, AddressProblem and DefendSolution.

Fig. 6 Referenced SuccessfulLearner Class Description in Protégé.

Perhaps, to address LQ2, we need to establish the object property assertion for Uncomplete Learners to be able to determine what attributes distinguishes such learners from the Successful ones. In view of that, Uncomplete Learner is a subclass of, amongst other NamedLearnerCategory, a Person that performs some Learning ActivityConcept who has a universal object property restriction/relationship with only some of the milestones of ResearchProcess Class but not all of the classes. As defined in Fig. 7 - the *necessary condition* is: if something is a Uncomplete Learner, it is *necessary* for it to be a participant of the Learning ActivityConcept class and *necessary* for it to have a kind of sufficiently defined condition and relationship with only some of the Class:- DefineTopicArea, ReviewLiterature, AddressProblem but not all four.

Fig. 7 Referenced UncompleteLearner Class Description in Protégé

The Object Property Restriction is used to infer anonymous classes that contains all of the individuals that satisfies the restriction, in essence, all of the individuals that have the relationship required to be a member of a particular Class. The consequence is the *necessary and sufficient Condition* which makes it possible to implement and check for consistency in the model, hence, it is necessary to fulfil the condition of the universal or existential Restriction - for any individual to become a member of a Class, as we answer using the LQ1 and LQ2 to describe the class SuccessfulLearner (Fig. 6) and UncompleteLearner (Fig. 7).

4. Semantic-based Algorithm Formalization

The Semantic Learning Process Mining (SLPM) algorithm describes the basis for our approach. To explain the strategies for constructing the classification of learning activity concepts and sub sets, we require the following notations; a, b, c, $d \in R$, is a function with domain R and process logs a, b, c, d. The domain R is a SuperClass of the SubClasses a, b, c, d. The SubClass (also referred to as SubSet) is a set where each individual Learning Activity occurs and sometimes may occur multiple times. For example, [$a1$, $a2$, $a3$, $a4$, $a2$, $a5$] may be the sequence set of learning activity for Person, $P \dots n$ over a, (the DefineTopicArea Milestone Class). i.e. $P \dots n(a) = |n \subseteq \mathcal{L}a|$.

Therefore, IF $a1$ = Define Topic event; $a2$ = Approval event; $a3$ = Topic decline; $a4$ = Refine Topic; $a5$ = End Topic Proposal;

THEN, the sequence set of activities for $P \dots n(a)$ = {*Define Topic event, Approval event, Topic Decline, Refine Topic, Approval event, End Topic Proposal*}.

However, for the purpose of the learning question LQ1 and LQ2, our focus is not only on the various individual activities that makes up a definitive Class (milestone) but on computing the set of individual process instances that has or not completed a given number of milestones. To complete a given milestone, one must perform the set (or perhaps a subset) of the activities that comprise it. Given the fact for transition purposes, a process instance does not move on to the next milestone without completing a distinctive sequence set of learning activities that makes up the milestone. The sum and difference in process logs for a given number of person, *P*, is defined in a straightforward way, i.e;

$$P \dots n = |n \subseteq \mathcal{L}a| \pm |n \subseteq \mathcal{L}b| \pm |n \subseteq \mathcal{L}c| \pm |n \subseteq \mathcal{L}d|.$$

Therefore, $P \dots n$ is a finite set $|n \subseteq \mathcal{L} \in R|$. For example, we described in Fig. 9 that Every Person that hasCompleteMilestone a DefineTopicArea and that hasCompleteMilestone a ReviewLiterature and that hasCompleteMilestone an AddressProblem and that hasCompleteMilestone a DefendSolution is a SuccessfulLearner". Therefore, the Class of Successful Learners, *PSL,* will be the sum of the set of activities log, $\mathcal{L}$, that a learner has completed for the milestones a, and b, and c, and d. Hence,

IF **PSL** is a Class that consist of the set $|SL \subseteq \mathcal{L}a| + |SL \subseteq \mathcal{L}b| + |SL \subseteq \mathcal{L}c| + |SL \subseteq \mathcal{L}d|$

THEN **PSL** is the set $|SL \subseteq \mathcal{L} \in R|$ as described in (Fig. 8)

Similarly, we defined in Fig 10. that Every Person that hasOnlyCompleteMilestone a DefineTopicArea or that hasOnlyCompleteMilestone a ReviewLiterature or that

hasOnlyCompleteMilestone an AddressProblem is an UncompleteLearner. Therefore, the Uncomplete Learners, **PUL,** is the class of leaners where some or set of activities for the milestone a, or b, or c, or d is missing over a finite set $|n \subseteq \mathcal{L} \in R|$. Hence,

IF **PUL** is a Class that consist of the set $|UL \subseteq \mathcal{L} \in R{-}a|$ or $|UL \subseteq \mathcal{L} \in R{-}b|$ or $|UL \subseteq \mathcal{L} \in R{-}c|$ or $|UL \subseteq \mathcal{L} \in R{-}d|$

THEN **PUL** is the set $|UL \subseteq \mathcal{L} \in R{-}1|$ as described in (Fig. 8)

SLPM Algorithm Formalization

Let $\mathcal{L}$, be a process log for *Person, P*, over *Research process, R*, and $a, b, c, d \in R$

Where: $a =$ DefineTopicArea Milestone
$b =$ ReviewLiterature Milestone
$c =$ AddressProblem Milestone
$d =$ DefendSolution Milestone

IF $P \ldots n$ is a measure of the number of times a, b, c, d occurs in R for *Person, P*

$$P \ldots n = |n \subseteq \mathcal{L} \in R|$$
$$P \ldots n = |n \subseteq \mathcal{L}a| \pm |n \subseteq \mathcal{L}b| \pm |n \subseteq \mathcal{L}c| \pm |n \subseteq \mathcal{L}d|$$

THEN
SuccessfulLearner, $PSL = |SL \subseteq \mathcal{L} \in R|$

$$PSL = |SL \subseteq \mathcal{L}a| + |SL \subseteq \mathcal{L}b| + |SL \subseteq \mathcal{L}c| + |SL \subseteq \mathcal{L}d|$$

UncompleteLearner, $PUL = |UL \subseteq \mathcal{L} \in R{-}1|$

$$PUL = |UL \subseteq \mathcal{L} \in R{-}a| \text{ or } |UL \subseteq \mathcal{L} \in R{-}b| \text{ or } |UL \subseteq \mathcal{L} \in R{-}c| \text{ or } |UL \subseteq \mathcal{L} \in R{-}d|$$
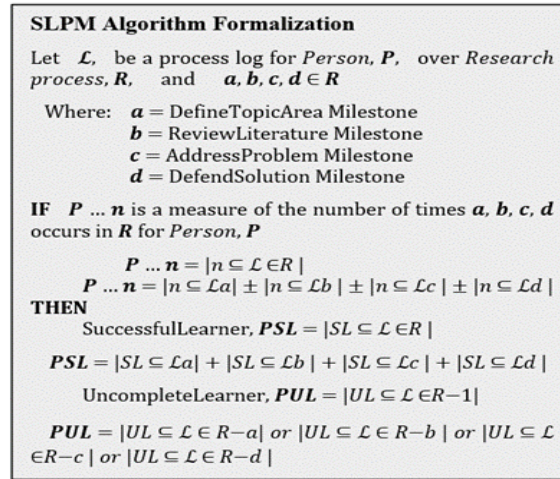
Fig. 8 The Semantic-based Learning Process Mining Algorithm formalization
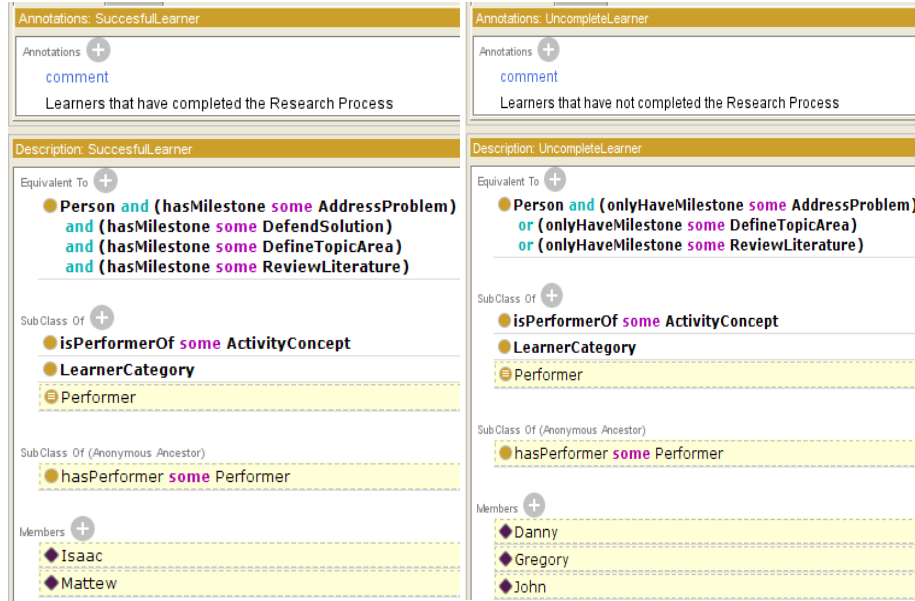

Fig. 9 Successful Learner Class   Fig. 10 Uncomplete Learner Class

5. Related Works

Vast number of information processing and retrieval systems in current literature use various mining techniques for representation of concepts, knowledge or data which are focused on applying technologies to different aspects of processes [7][8]. Researches in Semantic Web and Technologies has led to quite mature standards for assembling and modelling domain knowledge about any process [9]. Currently, Semantic Web Ontologies has become a fundamental tool for information extraction and knowledge processing by providing a structure for distribution of conceptual models about any given process. According to the survey in [10] a well-designed information retrieval or mining system should present results and discovered behaviours in a formal and structured format in the capacity of being interpreted as domain knowledge and to further enhance the existing knowledge base. The authors mention that ontology is one of the way to formally represent the mining results as sets of annotated terms and relations towards information extraction and association rule mining especially with Ontology-based Information Extraction (OBIE) systems [11]. Reference [10] also mention that ontology can integrate the use of heterogeneous/unrelated information to guide recommendation systems. According to the authors, ontology-based recommendation system uses ontology for user profiling and personalized search for data resources or patterns.

Elhebir and Abraham [12] notes that pattern discovery algorithms uses statistical and machine-learning techniques to build models that predicts behaviour of captured data. According to the authors, one of the most pattern discovery techniques used to extract knowledge from pre-processed data is Classification. They observe that most of the existing classification algorithms attains good performance for specific problems but are not robust enough for all kinds of discovery problems, and then propose that combination of multiple classifiers can be considered as a general solution for pattern discovery because they obtain better results compared to a single classifier as long as the components are independent or have diverse outputs. The approach compares the accuracy of ensemble models, which take advantage of groups of learners to yield better results using the Meta Classifier (Staking and Voting) alongside other Base classifiers: Decision Tree algorithm, k-Nearest Neighbour, Naive Bayesian and BayesNet. Explicitly, the works in [13][14][15] shows that the problems of modelling learning processes can be solved by transforming ontology population problem to a classification problem where, for each entity within the ontology, the concepts (classes) to which the entities belongs to have to be determined, hence, classified.

According to reference [16] Classification is one of the most common data mining technique that aims at finding models or functions that describes or distinguishes data classes/concepts. A useful application of such approach is to annotate the classification labels with the set of relations defined in an ontology especially for use in semantic enrichment of captured data. Semantics encoded in classification tasks has the potential not only to influence the labelled data but also to handle large number of unlabelled data [17][18]. The authors in [18] integrated ontology as consistency constraints into multiple related classification tasks by classifying multiple categories of unlabelled data in parallel to determine labels that violates the ontology. Reasoning on ontological knowledge plays an important role in the semantic representation of processes such as the learning process. This is possible because semantic reasoning allows the extraction

and conversion of explicit information into some implicit information, for instance, the intersection or union of classes, description of relationships, object properties and concepts assertions. In this paper, we apply the semantic-based approach to manage perspective of process mining. The focus is to further enhance this area of research by not only adapting the process mining tools but also present a way to relate semantic-based reasoning for computing relationships and ascertain concepts within a learning process domain by automatically constructing process models capable of defining, classifying and enhancing observed patterns or behaviours.

6. Conclusion and Future Work

The work in this paper proves that semantic-based process mining and analysis is a useful technique especially in solving some didactic issues and answering some questions with regards to different learning patterns/behaviour. We extract streams of event logs from a learning execution environment and then describe formats that allows for mining and improved analysis of the captured data. The approach makes use of semantic annotations and process description languages to link elements in event log of a research process with concepts that they represent in an ontology specifically designed for representing learning processes. By tackling the motivation of this paper, we delivered means by which the objectives and focus of the semantic approach contributes to the body of knowledge in current literatures. In summary, the main contributions of this paper are:

(1) Semantic motivated synchronization of event log formats for learning process data.
(2) Ontology driven search for explorative analysis of learning activities and execution.
(3) Techniques for annotating unlabelled learning activity sequences using ontology schema/vocabularies.
(4) Use of semantics tools to manage perspectives of process mining algorithms and definition of methods towards discovery and enhancement of process model analysis.
(5) Useful strategies towards development of process mining algorithms that are more intelligent, predictive and robotically adaptive.
(6) Importance of semantics process mining to augment information value of data about a domain process: case study of learning process.

Prospective researches could focus on adopting the approach described in this paper to help in analysing the streams of events logs that are involved in a different process domain, in order to produce inference knowledge that can be used to load a more enhanced model within the process domain area.

References

1. Kriegel, H.P., Borgwardt, K.M., Kröger, P., Pryakhin, A., Schuber, M., Zimek, A.: Future Trends in Data Mining. In *Journal of Data Mining and Knowledge Discovery.* Vol. 15, pp. 87-97 (2007)

2. Van der Aalst W.M.P.: Process Mining: Discovery, Conformance and Enhancement of Business Processes, Springer (2011)

3. de Medeiros, A.K.A, Pedrinaci, C., Van der Aalst, W.M.P, Domingue, J., Song, M., Rozinat, A., Norton, B., Cabral, L.: An outlook on semantic business process mining and monitoring.

In: *3rd Int. IFIP Workshop on Semantic Web & Web Semantics* (SWWS '07), 29-30 Nov 2007, Vilamoura, Portugal, pp. 1244–1255 (2007)

4. Disco Users Guide. Available at: https://fluxicon.com/disco/files/Disco-User-Guide.pdf.

5. Jareevongpiboon, W., Janecek, P.: Ontological approach to enhance results of business process mining and analysis. In *J. of Business Process Management,* Vol. 19(3) pp. 459-476 (2013)

6. Protégé Editor tool. http://protege.cim3.net

7. IEEE Task Force on Process Mining: Process Mining Manifesto. BPM Workshops LNBIP, vol. 99, pp. 169-194. Springer, (2012).

8. Fahland, D., Van der Aalst, W.M.P.: Repairing process models to reflect reality. In: *Business Process Management*, vol. 7481 of Lecture Notes in Computer Science, pp. 229-245. Springer, (2012)

9. Bemers-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. In *Scientific American*, Vol 284(5), pp 28-37, (2001).

10. Dou D., Wang, H., Liu, H.: Semantic Data Mining: A Survey of Ontology-based Approaches. In *Proc. of the IEEE 9th Int. Conference on Semantic Computing*, pp. 244 – 251, (2015).

11. Wimalasuriya, D.C, Dou, D.: Ontology-based information extraction: An introduction and a survey of current approaches. In *Journal of Info. Science*, Vol 36(3), pp 306-323, (2010).

12. Elhebir, M.H.A, Abraham, A.: A Novel Ensemble Approach to Enhance the Performance of Web Server Logs Classification. *Int. J. of Computer Information Systems and Industrial Management Applications* ISSN 2150-7988 Vol. 7 pp. 189-195, (2015).

13. d'Amato, C., Fanizzi, N., Esposito, F.: Query answering and ontology population: An inductive approach. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M., eds.: *Proc. of the 5th Euro. Semantic Web Conf.*, Vol. 5021 of LNCS, 288-302, (2008).

14. Okoye, K., Tawil, A.R.H., Naeem, U., Lamine, E.: A Semantic Reasoning Method Towards Ontological Model for Automated Learning Analysis. In *Advances in Nature and Biologically Inspired Computing*, Springer International Publishing, pp. 49-60, (2016).

15. Okoye, K., Tawil, A.R.H., Naeem, U., Lamine, E.: Discovery and Enhancement of Learning Model Analysis through Semantic Process Mining*. *Int. Journal of Computer Info. Systems and Industrial Management Applications*. ISSN 2150-7988 Volume 8, pp. 093-114, (2016).

16. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, US A, (2005).

17. Allahyari, M., Kochut, K.J., Janik, M.: Ontology-based text classification into dynamically defined topics. In *IEEE Int. Conf. on Semantic Computing* (ICSC), pages 273-278, (2014).

18. Balcan, N., Blum, A., Mansour, Y.: Exploiting ontology structures and unlabeled data for learning. In *Proc. of the 30th Int. Conference on Machine Learning*, pp. 1112-1120, (2013).