Machine Learning Based On Big Data

Extraction of Massive Educational Knowledge

https://doi.org/10.3991/ijet.v12.i11.7460

Abdelladim Hadioui(✉), Nour-eddine El Faddouli,
Yassine Benjelloun Touimi, and Samir Bennani
Mohammed V University in Rabat Agdal AV, Raba, Morocco
ahadioui@gmail.com

Abstract—A learning environment generates massive knowledge by means of the services provided in MOOCs. Such knowledge is produced via learning actor interactions. This result is a motivation for researchers to put forward solutions for big data usage, depending on learning analytics techniques as well as the big data techniques relating to the educational field. In this context, the present article unfolds a uniform model to facilitate the exploitation of the experiences produced by the interactions of the pedagogical actors. The aim of proposing the said model is to make a unified analysis of the massive data generated by learning actors. This model suggests making an initial pre-processing of the massive data produced in an e-learning system, and it's subsequently intends to produce machine learning, defined by rules of measures of actors knowledge relevance. All the processing stages of this model will be introduced in an algorithm that results in the production of learning actor knowledge tree.

Keywords—learning analytics, operational data, machine learning, big data analysis, knowledge management

1 Introduction

Currently, the field of education is flourishing rapidly throughout the world, due to the changes that have been occurring in this area with the implementation of Massive Open Online Courses MOOCs [1]. Great many research projects have been funded in order to draw the attention of researchers in this field to work on such massive data, conducting in-depth studies of MOOCs (COURSERA, OPEN ODX, etc.).

MOOCs generate big data in the form of activity traces. Such data are of three various types, namely: structured, semi-structured and unstructured data. In [2], the author has conducted an in-depth study on the types of data generated by the interactions of educational actors in online learning systems. The structured data are those found in the databases; the semi-structured are those found in the XML and JSON files, whereas the unstructured are those found in documents, video recordings, audio, etc.

Big data analysis [3] represents the combination of big data techniques with learning analysis. This combination enables to envision integrating Learning analytics (LA) algorithms with learning systems based on big data. Learning analytics (LA) represent a set of algorithms useful for the analysis and pre-processing of the massive data originally generated in the MOOCs. Indeed, we find two approaches: one supervised and another unsupervised [4,5]. On the other side, big data represent the tendency of actors to store massive data of different natures and to process them in parallel in tune with an architecture [6] built on three key elements: HDSF, MapReduce and YARN.

1.1 Research problematic

The massive data generated by the services which are offered within the MOOCs systems are structured, semi-structured and unstructured. Given such fact, prerequisite is to make an in-depth analysis focusing on all the massive data dimensions. To this end, the author in [7] identifies three dimensions of learning systems based on big data. A Learning system generates massive data that are:

- Varied: these are the types of data which are structured, semi-structured and unstructured. This constraint complicates the phases of knowledge extraction.
- Voluminous: these are big data that can reach TIRA Bit. Given this constraint, there is a large amount of data which are generated through the actor interactions.
- Distributed: these are massive data which are stored on multiple servers as well as different locations. It should be noted that the problem of knowledge distribution also constitutes a major constraint in the process of knowledge extraction.

With reference to the said constraints, in order to achieve our objective, we intend to suggest a solution to the use of the FRAMEWORK Mapreduce. This allows performing the parallel processing of massive data first, and then it allows the creation of a machine learning system based on the rules of measurement of relevance of the knowledge acquired in a learning system MOOCs.

The main objective of this paper is to propose a unified model for the analysis and pre-processing of actors experiences produced during the course of on-line learning activities, drawing on data mining algorithms and learning analytics techniques. Having criticized learning analytics models and approaches exist in the recent works and which deal with big data analysis problems in the area of education, the paper at hand proposes a deep analysis model buttressed on the interactions of pedagogical actors.

Indeed, actor experiences represent their productions in an on-line learning system. They represent the result of interactions in the system. Such interactions can be in structured, semi-structured (i.e., the log of operations of the actors traced by the system) and unstructured (i.e., documents, videos and audios, etc.). This results in a rich environment for the development of an actor profile. This production represents the input elements of the model proposed in this paper. This model will yield a tree of relevant knowledge of learning actors.

The scope of this article allows us to undertake learning analytics for a MOOCs platform existing in Morocco, integrating big data analytics techniques. Before de-

scribing this proposed model and the technologies necessary to implement it, it is of vital importance to examine the potentials of current standards and how they are to be used to create a learning system capable of generating massive knowledge.

2 State of art

The current learning system generates knowledge in the form of educational big data. Such knowledge is the consequence of the interactions undertaken by pedagogical actors in MOOCs. As a result of our in-depth analysis of the existing literature, which integrates techniques of analyzing massive data in education systems, we have noted that several outstanding research projects have been proposed by educational experts [8]. They focus on the handling of the educational actors experiences in a system based on big data, availing themselves of the learning analytics techniques for the extraction of massive knowledge relating to the distance learning field.

2.1 Massive data in the education field

The field of education witnessed evolution at all levels: (1) volume, (2) storage location, (3) the nature and type of massive data. This progress directly influences the pool of massive knowledge produced by learning actors in MOOCs, which explains the existence of many models and approaches in the literature that deals with massive data in a novel architecture based on massive data [9]. This is done with the aim of extracting the actor knowledge.

The author in [10] proposed state of art drawing on massive data produced by learning systems, notably the massive data of pedagogical field's which constitute the key elements of the present work. The author in this state of art has developed some methods dedicated to the creation of communication interfaces between a learning system and the new big data architecture. Next, he laid out some techniques and methods of learning analytics for such massive data. Many researchers have worked on the open source HADOOP Ecosystem [11], a system that processes massive data. The said architecture was proposed by the scientific community to give a large coverage of the massive data produced by learning actor interactions. HADOOP implementation aims at making use of its services in order to put at the disposal of educational actors the optimal methods for exploiting the massive data generated by actor interactions. A HADOOP system comprises the following elements: HDFS for the system of distributed files, MAPREDUCE for processing, HDFS files based on the functions that the MAPREDUCE and YARN Framework offers.

On the other hand, the author in [12] studied big data implementation in open education. This study arouses a considerable interest among researchers due to the changes in the way data are processed given the use of novel technology. The author also laid out the challenges of the implementation of learning systems that produce massive data for open education throughout the world, focusing on the contributions of MOOCs implementation based on big data. We find in learning system such produce a massive data, three dimensions studied by researchers. [13] These dimensions are

the key elements that have made researchers change their point of view and shift away from the traditional method to No-SQL, being mindful that the bulk of the massive data generated today is characteristic of an unstructured aspect. Therefore, we are forced to shift to the No-SQL method, which represents the mode of pre-processing the massive data generated by learning actors.

2.2 The analysis of learning for open education

Learning analytics integration into education has emerged since 2000. This technique involves several methods for the analysis and pre-processing of the massive data initially produced by actors in an e-learning system. These methods are useful for the creation of relevant knowledge owing to their bearing upon the initial massive data. Below, we present a comparative table of approaches to the learning analytics, proposed in the existing literature [30]:

Table 1. Learning analytics methods.

Learning Analytics Methods
Content analysis: The resources created by learning actors via their interactions
Discourse analysis: The analyses are based on the approaches of human sciences.
Analysis of social networks tools: That facilitates the interactions of educational actors.
Layout analysis: it seeks to understand the learner's dispositions towards his or her own learning, and how they relate to it.

In this connection, the author in [14] suggested some uses of leaning analysis methods in open education. He incorporated learning analytics techniques into massive data (originally created through the interactions of the actors involved in MOOCs) and applied several algorithms proposed by LA on them.

On the other hand, building on our analysis of the studies of the above-mentioned models, we found among those studies that in [15] the author made a statistical study of all the algorithms. Such a study indicates that all algorithms have borne upon the quality of data. Besides, the study proves that classification algorithms yield a better result in terms of massive data processing.

2.3 Analysis of massive data: technique and methods

The literature on big data based systems revealed several methods of massive data analysis. On the one hand, we found data mining methods that deal with structured and semi-structured data. For example, in [16] proposed a data mining model for learners knowledge extraction based on the best of Business Intelligence (BI) methodologies for dealing with massive data. The field of education has taken on an approach called EMD (Educational Data mining), which represents a range of methods and techniques of pre-treatment for educational big data. In the same vein, in his book, the author [17] made a review about the techniques and their contributions to develop educational actors profile, with view to propose cases of use of classification, clustering, association and decision trees applicable to open education systems. On

the other hand, we noted the existence of methods that deal with unstructured data such as videos, audios as well.

Worth noting, massive unstructured data represent a significant amount of data in learning systems that generate massive data. This claim was supported by the study conducted by the author [18].

The existing literature deals with a number of methods that tackle unstructured data. For example, the author in [19] proposed a classification model for unstructured documents via combining the Naïve Bayes classifier with the predictive analysis for learning systems. Also, the author in [31] proposed a model of digital educational resources indexing and dynamic user profile evolution. The aim of the model is to develop user profiles. However, most research efforts are still to be directed towards the operational data layer.

2.4 MOOCs (Massive Open Online Course)

MOOCs have been integrated into the field of education given the projects financed by numerous partners. Among those projects, we may refer to the MOOC Morocco [20]. Such projects are currently being implemented. They have been launched recently for the setting up of an e-learning platform devoted to educational actors. In order to implement such MOOCs, the author in [32] proposed to use of an analytical formalism to diagnose and evaluate MOOCs, using logical stepwise analytical approach.

A pedagogical actor generates knowledge through these interactions with online educational activities occurring in the platform. In [21], the author suggested an inventory of the activities performed by pedagogical actors. In his study, the author displayed that an actor produces knowledge through a number of activities. On the other hand, he noted that knowledge is stored in various locations and in the form of several structures such as videos, images, texts, etc. These data are produced in educational activities including: wiki, forums, lesson, etc.

Our analysis of the massive data generated by pedagogical actors in a MOOC permits to see that the MOOCs offer a significant wealth to educational actors thanks to its important digital data, something which was proved by the author in [22]. These massive data have been exploited by many researchers. For instance, the author in [23] suggested the usage of data mining for the exploitation of aforementioned experience traces. He laid out an approach premised on the classification of learners.

Given the advance of the MOOCs usage, we have encountered several problems that can be represented as research objects for experts in the field of open education. In this connection, many authors have worked on learning analytics methods in the field. This discipline is devoted to measuring, collecting, analyzing and reporting on e-learning processes.

Techniques and tools integration of learning analytics into MOOCs has been exploited in many studies. For instance, in [24] the authors proposed models and approaches for the analysis and pre-processing of massive data produced by learners, integrating data mining techniques into this domain. Other work [33], the author has proposed strategies for enhancing the learner experience and quality of MOOCs. Educational data mining offers a range of algorithms for the field of education.

Hypothesis

- Massive data generated by educational actors represent the best knowledge sources.
- Learner profiles development is based on the usage of data mining algorithms and learning analytics methods.
- In big data, unstructured knowledge represents the massive knowledge produced by an actor.
- The extraction of more than 50% of the traces generated by actors in the MOOCs represents a success factor for the e-learning system.

The massive knowledge produced by the actors of learning is useful for tutors in stages of the development of actor profiles.

3 Methodology

In our study, we need to make a profound analysis of the MOOCs in order to identify the core elements that can be extracted from the stored massive data of the pedagogical field. In doing so, we will undertake the following steps:

- Analyzing the nature and type of experiences produced by a pedagogical actor. In this phase, focus is to be lent to some existing learning systems such as OPEN EDX, COURSERA, MOODLE, etc. Then, we identify the massive knowledge generated by means of educational actors interactions. To make it clear, we will analyze the different structures of the massive data yielded by the learning actors experiences. These experiences represent all the knowledge that can be extracted from MOOCs. The experiences produced by pedagogical actors reflect the identified knowledge of their effects and productions in the platform, which are created in forums, wiki, tests, homework, audio / video productions ... etc.
- Studying some techniques of massive data processing through learning analytics and big data in order to underscore the best techniques and algorithms that will be useful in the present study context.
- Proposing our theoretical approach which is capable of solving the research problem and thus providing a theoretical main stay for the study; then, proposing a MapReduce model to make parallel processing of the massive data generated by the actors of learning drawing on the analysis of the massive data in an on-line learning system.

The following figure shows our methodological process:

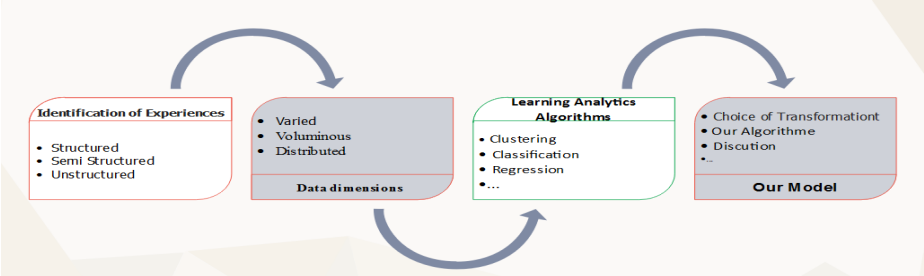

Fig. 1. The methodological process

For making the process in figure 1 succeed, we commit our efforts to a thorough analysis of the existing approaches, suggesting the model that solves the constraints of the approaches cited in the literature.

4 Contribution

This work proposes to tackle the layer of massive operational data. The said layer represents one of the key elements constituting the Conceptual Framework proposed in [25]. In the work at hand, we propose a uniform model for massive data analysis produced in MOOCs.

This model is concerned with operational massive data layer of our generic system, and it can be integrated into all educational systems.

The starting point for formulating the model is to cover the massive data produced by online learning systems; the learning systems platform implemented in Morocco is a case in point. Then, exploit massive data techniques as well as analysis of learning for the educational system. This exploitation technique is what makes the unstructured data, operational and exploitable by learning actors, offering more opportunities, more competitiveness for them in an online training session.

For this purpose, we will do more research on the strengths and flaws of the approaches cited at the beginning.

4.1 Theoretical ground of the MapReduce model

In this section, we lay out an environment that has - as input- massive, varied and voluminous data of different structures. β1, β2 and β3 are three sets of massive data: β1 represents the massive structured massive data, β2 has to do with the semi-structured and β3 contains the unstructured. Applying the MapReduces algorithms for these three sets, we have three semi structured files of learning actors knowledge, each file contain a population of knowledge.

- β1 is represented by the attributes: {@ActrorId, @Sp, @Level, @Knowledge};
- β2 is represented by the attributes: {@ActrorId, @Sp, @Level, @ Knowledge};

- β3 is represented by the attributes: {@ActrorId, @Sp, @Level, @Url Knowledge, @ Desc}.

The two attributes @Knowledge and @UrlKnowledge represent a set of sub-attributes of actors experiences in learning systems that produce the massive data of all structures (They represent all the productions of the actors in an MOOCs).

According to these definitions, and based on the work in [25], a framework of extraction of the knowledge of learning actors is proposed. It should be noted that this framework proposes two basic layers: the first one concerns the layer of operational massive data; whereas, the second has to do with the semantic layer for the representation of knowledge extracted by the use of domain ontology. This model deals with the big data of the operational data layer.

Next, we account for the representation of XML tree as follows:

We have the alphabet $\psi = \psi_{ele} \cup \psi_{att} \cup \psi_{data}$, where ψ_{ele} represents the set of element names, ψ_{att} represents the set of attribute names, and ψ_{data} is the set of data.

An XML document is represented by a triplet T = (t, type, value) like:

- The tree t represented by the function $t : Pos(t) \rightarrow \Psi \cup \{n\}$. For each position $p \in pos(t)$, $t(p) = a$ indicates that the symbol $a \in \Psi$ is associated with the position node p.
- The root is associated with the position ε and the empty tree is the tree $\{(\varepsilon, n)\}$ where $n \in / \Psi$ is a reserved symbol for the empty tree. This tree t is called the XML tree.
- The functions type and value are defined for any position $p \in$ post (t):

The type function: $t \times pos(t) \rightarrow \{data, element, attribute\}$ is defined :

$$type(p) = \begin{cases} data \text{ if } t(p) = \Psi_{data} \\ element \text{ if } t(p) \in \Psi_{ele} \\ attribute \text{ if } t(p) \in \Psi_{att} \end{cases}$$

The value function: $t \times Pos(t) \rightarrow Pos(t) \cup V$ is defined by: value (p) = p if type (p) = element; else value (p) = val ∈ V where V is a recursively innumerable set.

As though -t- is a tree, we can use the classical functions applied on trees: for every node of position p in pos (t), successor(p) is the set of successor nodes of the position p and predecessor(p) is predecessor node of p.

In this system, we have as input element the set β which includes three subsets defined as follows:

- β1 the set of structured data;
- β2 the set of semi-structured data;
- β3 the set of unstructured data.

Each set is represented by elements defined by the attribute / values pair.

Then we insert the massive data considered relevant for our XML tree. The insertions are extracted from the interactions of learning actors in order to automate the operational data layer of our Conceptual Framework [25] by relevant experiments extracted from the MOOCs learning system.

In what follows, we integrate the evaluation criteria of the actors, it should be noted that these criteria have been proposed in our model proposed for qualifying learning actors knowledge [26]. Such a model integrates the criteria of evaluation in the form of a test for assessing the actor level according to a scale of 1 to 5 for the criteria taken to be necessary for the positioning of the systems users (Intelligence, memory, style, etc.). The relevance matrix is defined as follows:

$$M = \sum_{i=1,j=1}^{m,n} Mij \tag{1}$$

Additionally, we exploit the MapReduce functions for the good functioning of our target system given the contributions provided by this Framework at the level of massive data extraction. In this sense, we have two functions mapreduce: the map function and the reduce function, which are defined as follows:

- The map function => {k, v} ;
- The reduce function => list {k, v}.

In this context, these functions are used to make parallel processing of the knowledge acquired by the processing of the massive data produced by pedagogical actors in order to process files containing structured, semi-structured and unstructured massive data.

4.2 The Proposed Algorithm

INPUT:

βi,j represents the set of knowledge categories: structured, semi-structured and unstructured as follows:

$$\beta = \sum_{i=1,j=1}^{m,n} \beta ij ,$$

β1 is the structured knowledge;
β2 is semi structured knowledge;
β3 is unstructured knowledge.

$M = \sum_{i=1,j=1}^{m,n} Mij$ / is the set of learning actors evaluation criteria, i represents the number of lines and j represents the number of attributes.

1. The merging of knowledge sets with the set of actors evaluation criteria produce the set π,

   ```
   π = Merge (M, β).
   ```

2. The fractionation of the input files whatever their structures, defined by the set II

   ```
   Split(π) In subsets πi, such that i from 1to n
   ```

3. Path of the subset / for i from 1 to m applied the MAP function to the already split files.

   ```
   MAP (πi; ki, vi).
   ```

4. The identification of the sets of elements that consist of classes => elements: attr, val.
5. Adding the node to the global tree:

   ```
   Tree T = T.append (node).
   ```

6. `if i = m exit.`
7. Marching learning step: in this step this model puts forward a machine learning system created via rules that make the grouping of knowledge extracted from a learning system.

   We define the rules as follows:

- Rule 1: R 1 / empty knowledge attribute;
- Rule 2: R 2 / unknown actors attribute;
- Rule 3: R 3 / Repetitive Knowledge;
- Rule 4: R 4 / summation of the values of the evaluation criteria of learning actors attribute is bigger than 9 the knowledge are irrelevant;

Rule 5: R 5 / summation of the values of the evaluation criteria of the learning actors attribute less than 9 the knowledge are relevant.

With the implementation of the rules suggested by this system, three categories of knowledge are generated: unnecessary, irrelevant, and relevant.

8. Finally, this model is based on machine learning which will create three output trees according to the three actors categories proposed in this article.

OUTPUT: Classification of actor knowledge in accordance with the three categories.

4.3 MapReduce model

In this model, we rest on the mapreduce functions to extract knowledge from the educational actors in a learning system based on big data. The aim of this model is to split the inputs files created in any educational system, transforming such data into a

key / value pair. Such a splitting will be dealt with through several steps. This will enable us to make an iterative MapReduce to create an XML tree.

This model is designed for MOOCs learning system based on big data. The massive data generated by learning actors are those results emanate from the interactions of the actors in an e-learning system. Our analysis of these massive data revealed that they are either: structured, semi-structured or unstructured. Structured and semi-structured data are defined by the same attributes; whereas, unstructured data replaced the @knowledge attribute with @UrlKnowledge.

The current model proposes a solution for the extraction of learning actors knowledge based on the massive data generated in the MOOCs. This is shown in Figure 2.

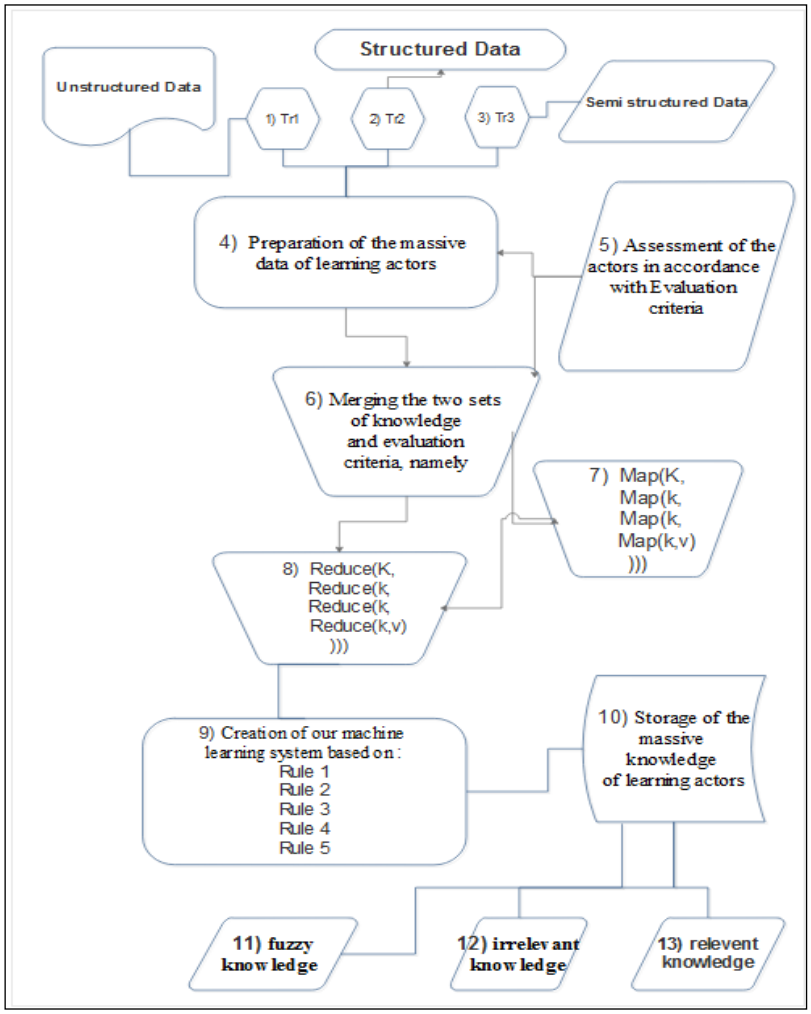

Fig. 2. MapReduce model of knowledge extraction in educational big data

This model consists of several operations, in the form of processes related to each other. These operations can be presented by the steps described as follows:

1) Pre-processing of unstructured massive data: In this phase, their initial processing is considered, proposing a metadata of unstructured files such as videos, audios, images, etc. These metadata are extracted on the basis of system logs in MOOCs online learning system.

These data are introduced by the set β3 = {@ActrorId, @Sp, @Level, @UrlKnowledge, @ Desc}. This set represents unstructured knowledge of learning actors, which is produced by the interactions of the learning actors.

2) And 3) Pre-processing of structured and semi-structured data: In this phase, we engage in the initial processing of the structured and semi-structured massive data. The result falls into two sets β1, β2 which are composed of the following attributes: β1, and β2 {@ActorId, @Sp, @Level, @Knowledge}.These two sets can be extracted from MOOCs learning system.

4) Preparation of the massive data of learning actors. In this phase we, make the grouping of the three sets cited above in terms of pre-processing of the data input. The aim of this operation is to integrate the three sets in the same engine of massive data grouping for creating a source of the massive knowledge produced via the interactions of learning actors.

5) Assessment of the actors in accordance with Evaluation criteria: In this operation we integrate the evaluation criteria of learning actors in order to give more relevance to their knowledge. These criteria were proposed by the work in [26]. We integrate the map() function to split our structured file already extracted from our MOOCs. The knowledge generated by the actors is the result of interactions in the national network.

6) Merging the two sets of massive knowledge and evaluation criteria, namely: the knowledge and all the evaluation criteria of the actors involved in learning system. In this model, we merge the sets received as input. Subsequently, this model proposes a massive knowledge associated with actors evaluation criteria. Moreover, the model performs several classifications and sorting in accordance with the classification rules proposed by data mining models.

7) Division of Input files: it has to do with the massive knowledge in which we have made the massive data grouping into the inputs of the mapreduce functions in order to be implemented for the proper functioning of this model. In this operation we will make an iterative map for the input sets so that this model makes the creation of a tree of the actors knowledge. The number of iterations is defined at the beginning thanks to the good functioning in our generic model.

8) Application of reduce function: The model in this stage integrates the second reduce function of the mapreduce Framework. The inputs of this function are the massive knowledge associated with the evaluation criteria of learning actors.

9) Creation of our machine learning system: In this phase, the model integrates a machine learning approach, which consists of several rules, in order to classify the massive knowledge of learning actors. Having integrated the proposed rules at the beginning of this work, we have three categories of actors knowledge: 1) the non-usable knowledge, that is, the empty knowledge; 2) unclear knowledge; 3) irrelevant

knowledge, that is to say, the knowledge of which the sum of the evaluation criteria of the actors is bigger than 9: 4). On the other hand, the relevant knowledge is the knowledge of which the sum of the evaluation criteria of the actors is less than equal 9.

Through the integration of the set of steps created in the beginning, we have an adequate system proposing knowledge relevant for learning actors.

10) Storage of the massive knowledge of learning actors: In this phase, the system generates an XML tree that yields the structuration of massive knowledge. These are produced by means of the interactions of learning actors while browsing through all the mentioned stages. The structure of our tree is proposed in our research [25], suggesting a Framework for the extraction of learning actors knowledge while following the layer of the operational data as well as the layer of the semantic data, using domain ontology for representing the massive knowledge based on the integrations of the actors.

5 Discussion

5.1 Results

By analyzing the results obtained via this model, we have, as a result, achieve XML tree that proposes the massive knowledge extracted from the system. The machine learning model proposed in this paper puts forward knowledge tree of learning systems by applying the knowledge insertions detected by our system with the aim of structuring the unstructured knowledge. In this sense, Figure 3 presents the relevant knowledge tree as follows:

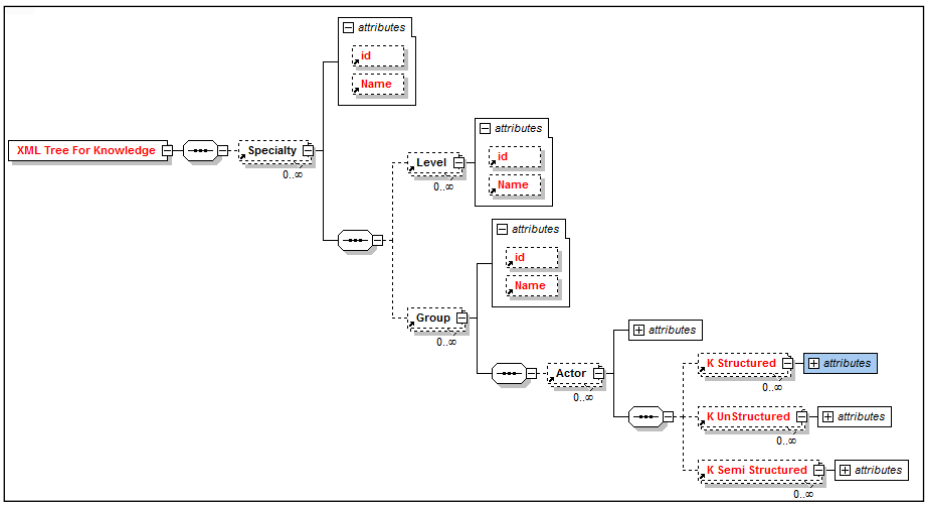

Fig. 3. Knowledge tree of learning actors created via the model proposed in this article

Figure 3: presents the prototype tree of learning actors, knowledge, which is created by the model proposed in this article. This figure shows a prototype of the tree to be created by machine learning system put forward by this research. In this prototype, the system creates an XML tree of the learning actor knowledge of all structures and proposes a knowledge tree in a semi-structured format.

In the article at hand, learning model proposes three categories of learning actors knowledge, based on the proposed rules. The prototype of the XML tree is created and enriched by the algorithm of the machine learning model proposed in this work.

The knowledge produced in such XML tree is the result of learning actors interactions in the MOOCs. These interactive experiences have been produced in the form of forums, wiki, audio, videos, etc. Subsequently, several operations have been proposed in the algorithm cited earlier. This model has been generalized via the processing of educational massive data in which data are distributed on several nodes. The approach adopted in this article generates efficient results due to the level of complexity proposed by aforementioned algorithm.

5.2 Results contribution

Through analyzing the results this model came up with, this article proposes an actor knowledge in the format semi-structured. This model deals with operational, massive data on open education. These data are in most cases represented in unstructured format; they are distributed over several nodes in accordance with big data architecture. This proposition, as well as the results obtained makes clear that we have an adequate model. This is due to the speed of these transactions which are the results of the initial processing carried out on massive data originally produced through actor interactions. This model proposes knowledge in the form of three categories:

Category 1 and 2: these are for fuzzy or irrelevant knowledge. This is an analytical factor for tutors in order to analyze the motivation of learners in online training.

Category 3: for relevant knowledge which represents the knowledge to be extracted to enrich the recommendation phase in order to give more autonomy to learning actors in phases of development of their profiles.

In so doing, this model pre-processes the result of the relevant knowledge available for the ontology layer.

Framework in [25] is automated due to the mapping techniques between operational data layer and semantic layer existing in the market [27]. Through this, we can subsequently measure the proper functioning of such framework which will be a good environment of relevant actor knowledge. In the future we can conduct a number of works on the knowledge extracted via learning systems, in which will be a possible recommendation of the relevant knowledge.

6 Conclusion And Future Work

The research subject matter of this work is the proposal of a novel model of actors knowledge extraction from the massive data generated in MOOCs. It lends heavy

focus to the actors experiences produced as a result of their interactions in e-learning system. Then, it puts forward a scenario of pre-processing of educational data in such a system. Besides, it gives an overview of the integration of learning analytics models into the phases of pre-processing of massive data in MOOCs, and it highlights the advantages of integrating massive data into the process of developing learner profiles.

This model results in knowledge sorts of learning actors, which are created in the form of three categories: unused knowledge; Irrelevant knowledge; relevant Knowledge.

The suggested model could be subject to subsequent experiments in order to measure its adequacy, and its results could constitute a starting point for the development of the ontology layer.

7 References

[1] Egloffstein, M., Ifenthaler, D., (2017). Employee Perspectives on MOOCs for Workplace Learning. TechTrends 61, 65–70. https://doi.org/10.1007/s11528-016-0127-3

[2] Gautier, J.-M., Gayet, A.,(2017). De la Data aux Big Data: enjeux pour le Marketing client–Illustration à EDF. Statistique et Société 4, 49–53.

[3] Wamba, S.F., Gunasekaran, A., Akter, S., Ren, S.J., Dubey, R., Childe, S.J., (2017). Big data analytics and firm performance: Effects of dynamic capabilities. Journal of Business Research 70, 356–365. https://doi.org/10.1016/j.jbusres.2016.08.009

[4] Liu, X., Wang, J., Yin, M., Edwards, B., Xu, P., (2017). Supervised learning of sparse context reconstruction coefficients for data representation and classification. Neural computing and applications 28, 135–143. https://doi.org/10.1007/s00521-015-2042-5

[5] Hu, Y., Baraldi, P., Di Maio, F., Zio, E., (2017). A Systematic Semi-Supervised Self-adaptable Fault Diagnostics approach in an evolving environment. Mechanical Systems and Signal Processing 88, 413–427. https://doi.org/10.1016/j.ymssp.2016.11.004

[6] Nkenyereye, L., Jang, J.-W., (2015). A study of big data solution using hadoop to process connected vehicle's diagnostics data, in: Information Science and Applications. Springer, pp. 697–704. https://doi.org/10.1007/978-3-662-46578-3_82

[7] Wang, L., Zhan, J., Luo, C., Zhu, Y., Yang, Q., He, Y., Gao, W., Jia, Z., Shi, Y., Zhang, S., others, (2014). Bigdatabench: A big data benchmark suite from internet services, in: High Performance Computer Architecture (HPCA), 2014 IEEE 20th International Symposium on. IEEE, pp. 488–499.

[8] Dietze, S., Taibi, D., d'Aquin, M., (2017). Facilitating scientometrics in learning analytics and educational data mining–the LAK dataset. Semantic Web 8, 395–403. https://doi.org/10.3233/SW-150201

[9] Gibson, D.C., Ifenthaler, D., (2017). Preparing the next generation of education researchers for big data in higher education, in: Big Data and Learning Analytics in Higher Education. Springer, pp. 29–42. https://doi.org/10.1007/978-3-319-06520-5_4

[10] Daniel, B.K., (2017). Overview of Big Data and Analytics in Higher Education, in: Kei Daniel, B. (Ed.), Big Data and Learning Analytics in Higher Education: Current Theory and Practice. Springer International Publishing, Cham, pp. 1–4. https://doi.org/10.1007/978-3-319-06520-5_1

[11] Sakr, S., (2016). Big Data 2.0 Processing Systems: A Survey. SpringerBriefs in computer science.

[12] Zhou, L., Pan, S., Wang, J., Vasilakos, A.V., (2017). Machine Learning on Big Data: Opportunities and Challenges. Neurocomputing.
[13] Bechini, A., Marcelloni, F., Segatori, A., 2016. A MapReduce solution for associative classification of big data. Information Sciences 332, 33–55. https://doi.org/10.1016/j.ins.2015.10.041
[14] Kache, F., Kache, F., Seuring, S., Seuring, S., (2017). Challenges and opportunities of digital information at the intersection of Big Data Analytics and supply chain management. International Journal of Operations & Production Management 37, 10–36. https://doi.org/10.1108/IJOPM-02-2015-0078
[15] Groves, P., Kayyali, B., Knott, D., Kuiken, S.V., (2016). The'big data'revolution in healthcare: Accelerating value and innovation.
[16] Khalil, M., Ebner, M., (2016). What is Learning Analytics about? A Survey of Different Methods Used in 2013-2015. arXiv preprint arXiv:1606.02878.
[17] Sancho, J., (2016). Learning Opportunities for Mass Collaboration Projects Through Learning Analytics: a Case Study. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje 11, 148–158. https://doi.org/10.1109/RITA.2016.2589482
[18] Pasichnyk, V., Shestakevych, T., '2017). The Model of Data Analysis of the Psychophysiological Survey Results, in: Advances in Intelligent Systems and Computing. Springer, pp. 271–281. https://doi.org/10.1007/978-3-319-45991-2_18
[19] Olson, D.L., Wu, D., (2017). Predictive Models and Big Data, in: Predictive Data Mining Models. Springer, pp. 95–97. https://doi.org/10.1007/978-981-10-2543-3_8
[20] Gandomi, A., Haider, M., (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35, 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
[21] Yuan, L., Powell, S., CETIS, J., others, (2013). MOOCs and open education: Implications for higher education.
[22] Michalik, P., Stofa, J., Zolotova, I., (2014). Concept definition for Big Data architecture in the education system, in: Applied Machine Intelligence and Informatics (SAMI), 2014 IEEE 12th International Symposium on. IEEE, pp. 331–334. https://doi.org/10.1109/SAMI.2014.6822433
[23] Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y., Herawan, T., (2014). Big data clustering: a review, in: International Conference on Computational Science and Its Applications. Springer, pp. 707–720. https://doi.org/10.1007/978-3-319-09156-3_49
[24] Kansal, N., Solanki, V.K., Kansal, V., (2016). Educational Data Mining and Indian Technical Education System: A Review. Feature Detectors and Motion Detection in Video Processing 18.
[25] Hadioui, A., Bennani, S., Idrissi, M.K., others, (2016). AN ONTOLOGICAL EXTRACTION FRAMEWORK OF THE ACTORS'PEDAGOGICAL KNOWLEDGE. Journal of Theoretical and Applied Information Technology 93, 69.
[26] Hadioui, A., Bennani, S., Khalidi, M., Faddouli, E., n.d. (2015). Modèle de Qualification des connaissances pédagogiques dans un système E-learning.
[27] Manikandan, S.G., Ravi, S., (2014). Big data analysis using Apache Hadoop, in: IT Convergence and Security (ICITCS), 2014 International Conference on. IEEE, pp. 1–4. https://doi.org/10.1109/ICITCS.2014.7021746
[28] Jadhav, A., Pandita, A., Pawar, A., Singh, V., (2016). Classification of Unstructured Data Using Naïve Bayes Classifier and Predictive Anaysis for RTI Application. An International Journal of Engineering & Technology 3.

[29] Kidziński, \Lukasz, Giannakos, M., Sampson, D.G., Dillenbourg, P., (2016). A tutorial on machine learning in educational science, in: State-of-the-Art and Future Directions of Smart Learning. Springer, pp. 453–459. https://doi.org/10.1007/978-981-287-868-7_54

[30] Nunn, S., Avella, J.T., Kanai, T., Kebritchi, M., 2016. Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. Online Learning 20.

[31] Slimani, H., El Faddoul, N., Bennani, S., Amrous, N., 2016. Models of Digital Educational Resources Indexing and Dynamic User Profile Evolution. International Journal of Emerging Technologies in Learning 11.

[32] Tahiri, J.S., Bennani, S., Idrissi, M.K., 2015. Using an analytical formalism to diagnostic and evaluate Massive Open Online Courses, in: Intelligent Systems: Theories and Applications (SITA), 2015 10th International Conference on. IEEE, pp. 1–6. https://doi.org/10.1109/SITA.2015.7358389

[33] Conole, G.G., 2015. MOOCs as disruptive technologies: strategies for enhancing the learner experience and quality of MOOCs. Revista de Educación a Distancia.

8 Authors

Abdelladim Hadioui, Nour-eddine El Faddouli, Yassine Benjelloun Touimi, and **Samir Bennani** are with RIME TEAM-Networking, Modeling and e-Learning- LRIE Laboratory-Research in Computer Science and Education Laboratory at Mohammadia School Engineers (EMI) - Mohammed V University in Rabat Agdal AV. Ibn Sina Agdal Rabat BP. 765 Morocco.

Article submitted 23 July 2017. Published as resubmitted by the authors 10 September 2017.