Adaptive e-learning using Genetic Algorithm and Sentiments Analysis in a Big Data System

Youness MADANI*
and Jamaa BENGOURRAM
GI laboratory
Departement of Computer Sciences
Faculty of Sciences and Technics
Sultan Moulay Slimane University
*Corresponding author

Mohammed ERRITALI
and Badr HSSINA
TIAD laboratory
Departement of Computer Sciences
Faculty of Sciences and Technics
Sultan Moulay Slimane University

Marouane Birjali
LAROSERI Laboratory
Department of Computer Sciences
Faculty of Sciences, El Jadida
University of Chouaib Doukkali

***Abstract*—In this article we describe our adaptive e-learning system, which allows the learner to take courses adapted to his profile and to the pedagogical objectives set by the teacher, we use for adaptation the genetic algorithms to give the learner the concepts that must learn in an optimal way by seeking the objectives most adapted to his profile. And after a second level of adaptation using one of the social networks of the learner (twitter, facebook, Google + ...), based on his post on one of these social networks we propose two levels of analysis. The first one is to look for the period of activity which gives us an idea about the period when the learner is active and the second consists of making an analysis of the feelings on the publications that are published during the period of activity and related to education. Our work therefore is to adapt the profile of the learner with the pedagogical objectives by using the genetic algorithm and the notions of the research of information by doing this work in a Big Data system, that is to say we parallelize the search problem using Hadoop with Hadoop distributed file system (HDFS) and the MapReduce programming model,and after using information from a social network of the learner, we look for the period of activity of the learner and the feeling (sentiment analysis) related to the publications of the period of activity.**

***Keywords*—Adaptive E-learning; genetic algorithms; research of information; social network; period of activity; sentiment analysis; parallelize the search problem; big data; Hadoop; MapReduce; Hadoop distributed file system (HDFS)**

I. Introduction

The new information and communication technologies (ICT) profoundly improve our ways of informing, communicating and training us. This technological emergence has revealed a new mode of learning known as e-learning. It is based on access to online training, interactive and sometimes personalized, distributed via a network (Internet or Intranet) or another electronic medium. This access makes it possible to develop the skills of the learners while making the learning process independent of time and place.

The field of research in e-learning is very broad. It is also the object of a prosperous industrial activity and e-learning research issues Could be described as questions about adapting educational practices with today's technology [1].

Since the beginnings of e-learning, Artificial Intelligence (AI) techniques have been tested to increase the learning experience. The use of AI algorithms in e-learning gives birth to adaptive e-learning.

Adaptive learning aims to propose a learning method that adapts to each learner's profile. This teaching emerged in the 1970 but is gaining momentum as technologies become more powerful and less costly. We are close to the research on artificial intelligence which has made enormous advances [1].

The goal of adaptive e-learning is to give the user of an e-learning platform a pedagogical content customized according to his profile with the use of algorithms like that of artificial intelligence. The idea is to find the pedagogical objective most adapted to the learner's profile because we can find a course with several pedagogical objectives suggested by a teacher.

We find several algorithms used in the literature in adaptive e-learning such as genetic algorithms that transform work into an optimization problem and give the learner the concepts that must learn.

Our work lies in this research axis, an adaptive e-learning system based on genetic algorithms and the notions of information retrieval like the similarity to find the relevant documents for a user request that expresses the user need.As a comparison between our work and information retrieval systems the learner profile will play the role of the query, and the different learning objectives will play the role of the sought documents,so it is as we want to calculate the similarity between the learner profile and the learning objectives to find those that are relevant to his profile and to keep only those objectives that are relevant that will then play the role of the initial population for the genetic algorithm in order to find an optimal objective, without forgetting that this work will be carried out in a parallel way by working with the Hadoop framework, storing the input and the output data of our algorithm in HDFS and using the Hadoop MapReduce programming model.

In addition to searching for an optimal pedagogical content for the learner's profile our work also consists in using one of the most powerful platforms in the world, it is the social networks, our proposal is to give the learner the possibility to connect to our e-learning platform using a social network like facebook, twitter, etc.The idea is to calculate a measure that we call it **period of activity** that will give us an idea about the period of the day when the learner is very active and therefore

will help us to improve the learner's skills, and after making a sentiment classification (motive,demotivate and neutral) of the publications of this period of activity that are related to the field of education to know the feeling of the learner that will help the teacher to define the level of learning.

The reminder paper is organized as follows: In Section 2 we give a brief overview of the Genetic Algorithm(GA) and Information Retrieval System(IRS). Section 3 gives some related work, and in Section 4 we explain how we used social networks in our work. A detailed description of our work is presented and detailed in Section 5,and in Sections 6 and 7 we describe/give some experimental results the parallelization's steps, the Section 8 summarize our work; finally, we give a conclusion and some future works in Section 9.

II. Genetic Algorithm and Information Retrieval System

A. Information Retrieval System

The search for information tries to solve the following problem: Given a very large collection of objects (mostly documents), find those that respond to a need for information expressed by a user (request). In the Information Retrieval System, we find a request and we want to find the objects (documents) that are relevant to it, the way to evaluate a document if it is relevant or not is to calculate the similarity between the request and that document.

Before the calculation of the similarity it is important to index all the documents and also the request that is to make them in a presentation to facilitate its use in our case we use the vector representation [2], where each element of the vector represents the weight (frequency) of each term or concepts in the document or in the query.

Our corpus in our case contains the documents that represent the learner's objectives, the first thing to do is to extract all the terms or concepts in the corpus, and for each document construct a vector That represents it, if a term exists in the document we calculate its weight and if not we put 0, at the end of this operation we construct a vector for each document to calculate the similarity between the profile of the learner and each pedagogical objective.

The calculation of the weights of terms or concepts in each document is calculated by the following formulas:

$$Poid(t_i, d_j) = TF * IDF \tag{1}$$

avec:

- $$TF = \frac{f(t_i, d_j)}{N} \tag{2}$$

  $f(t_i, d_j)$ is the number of occurrences of the term t_i in the document d_j and N is the total number of terms in the document d_j.

- $$IDF = \frac{log(f(t_i, d_j))}{M} \tag{3}$$

  $f(t_i, d_j)$ Is the number of occurrences of the term t_i in the document d_j and M is the total number of documents in the corpus.

The similarity used in our work is the Cosine similarity. This measure uses the complete vector representation, i.e. the frequency of the objects (words).Two objects (documents) are similar if their vectors are confused, the formula is defined by the ratio of the scalar product of the vectors X and Y and the product of the norm of X and Y.

$$Sim_{cos}(X, Y) = \frac{\sum_{i=1}^{n} x.y}{\sqrt{(\sum_{i=1}^{n} x^2)}.\sqrt{(\sum_{i=1}^{n} y^2)}} \tag{4}$$

B. Genetic Algorithm

Genetic algorithms (GAs) are stochastic optimization algorithms based on the mechanisms of Natural selection and genetics, their operation is extremely simple, we leave with a population of potential solutions (chromosomes) initial selected arbitrarily, we evaluate their relative performance (fitness). On the basis of these performances, a new population of potential solutions is created using simple evolutionary operators such as selection, crossing and mutation. This cycle is repeated until a satisfactory solution is found [3].

In our work we use a simple GA, which consists of iterating the following three operations: reproduction, crossing and mutation, the population created during each iteration is called a generation and it's noted P_t.

There has been an increasing interest in the application of GA tools to IR in the last few years. Concretely, the machine learning paradigm, whose aim is the design of a system able to automatically acquire knowledge by themselves, seems to be interesting on this topic.

The first thing in a genetic algorithm is the definition of the initial population (selection operator or evaluation) on which we will apply the treatment as in our case it is to show the documents (educational objectives) relevant to the profile of the learner using the cosine similarity that will play the role of fitness function which is a very important parameter in GA because with it we can decide whether an individual is going to be selected or not. There is a lot of methods to make the selection like the biased lottery, the elitist method or the selection by tournaments.

After applying the selection operator to the initial population, it is the reproduction step with the application of the crossing or crossover operation and the mutation operation.

In the literature we find many works that applies genetic algorithms in the search for information, as in [4] where authors use in their information retrieval system the genetic algorithm to find the relevant documents for a user query,they use the vector representation to present the documents of the search base and the query, they have made comparisons with precision measurements and recall of the system using different fitness functions like Cosine, Dice and Jaccard.

In [5] the researcher explored the problems embedded in this process, attempted to find solutions such as the way of choosing mutation probability and fitness function, and chose

Cranfield English Corpus test collection on mathematics. Such collection was conducted by Cyrial Cleverdon and used at the University of Cranfield in 1960 containing 1400 documents, and 225 queries for simulation purposes. The researcher also used cosine similarity and jaccards to compute the similarity between the query and documents and used two proposed adaptive fitness function, mutation operators as well as an adaptive crossover. The process aimed at evaluating the effectiveness of results according to the measures of precision and recall.

Vajitoru [6] Also uses the Genetic Algorithms in the research of information, he proposed a new operation of crossing to improve the research with the genetic algorithm, for that he made a comparison between his proposal and a classic GA and the results show the Effectiveness of its proposal.

Sathya and Simon [7] use the genetic algorithms to improve an information retrieval system and make it effective for obtaining more pages relevant to the user's query and optimize the search time.

In [8] the Researchers present a new fitness function for approximate information retrieval which is very fast and very flexible than cosine similarity.

Fan et al. propose an algorithm for indexing function learning based on GA, whose aim is to obtain an indexing function for the key term weighting of a documentary collection to improve the IR process [9].

III. Genetic Algorithm on Adaptive e-learning

Several works of artificial intelligence are used in adaptive e-learning to give the learner a content adequate to his profile in the literature we find:

Hawkes and Derry [10] have used the informal fuzzy reasoning in the TAPS system to determine with uncertainty the solution that the student has built among those of the system (models).

Ruiz et al. [11] have modeled an adaptive hypermedia system, called Feijjo.net, based on the learning style. The system uses fuzzy logic to determine the learner's style from the CHAEA questionnaire.

Chrysafiadi and Virvou [12] have proposed a learner model that represents the learner's knowledge through the overlay model (presented concepts that the learner master with "1" or with the word "known" and those that do not master by "0" or unknown), the fuzzy logic allowed to define and update the level of knowledge of each concept, with each interaction with the e-learning system.

Martin and VanLehn [13] have presented OLAE as an assessment tool that collects data from students solving physics problems in college. For each problem, OLAE automatically creates a Bayesian network that calculates the probabilities indicating the rules that the student uses.

Viccari et al. [14] have introduced AMPLIA, an intelligent learning environment used as a training tool in the medical field, the system combines bayesian networks with cognitive.

There are also works that use genetic algorithms for adaptive e-learning, namely:

the work of Romero et al. [15] Which represent a methodology to improve education systems, using grammar based on genetic algorithm techniques and multi-objective optimization to extract prediction rules allowing teachers to select the most appropriate changes to improve the efficiency of the Training.

Chang and Ke [16] Have proposed a customized composition of courses in an adaptive learning system, based on the genetic algorithm (GA), with the aim of specifying the appropriate learning resources for each learner.

In [1] the Researchers describe an adaptive system conceived in order to generate pedagogical paths which are adapted to the learner profile and to the current formation pedagogical objective. They have studied the problem as an "Optimization Problem" using Genetic Algorithms, the system seeks an optimal path starting from the learner profile to the pedagogic objective passing by intermediate courses to prepare the courses for adaptation.

In [17] a genetic algorithm based adaptive learning scheme for context aware e-learning has been described, the Researchers defined a new three level structure for learner's context comprising of the content level, presentation level and media level is defined. The learning path generation algorithm now evolves into a learning scheme generation as it generates a learning path accommodating the entire learner's context.

IV. Social Network

As we said earlier our adaptive e-learning system uses social networks to extract information about the learner for that we use social authentication.

Social authentication with a social network is an authentication type that allows us to use existing login information of a user to a social network such as Facebook, Twitter or Google+, to connect the user to a third website, instead of creating a new login account specifically for this site. Social login is simple and effective they allow users to authenticate to websites without having to create an additional account, Just a click on a social button authentification is Enough [18].

The authentification button increases the enrollment to a platform. Why? Just because the authentication button removes the need for the user to refill a form, choose a username and secure password. Thus, it now needs to do one click to move from one social network to another.

This module simplifies the registration of a new user to a site. Instead of filling the required fields for registration or login, you can simply click on the button corresponding to the social network, and that's all you are a registered user. With this module the number of registered users on your website increases as well as the potential activity of users.

We use the social authentication in our work to retrieve the learner's publications to analyze them afterwards either to look for the period of activity or to analyze the feelings of the publications that are in relation with education.

A. Period of Activity

The period of activity is a measure proposed in our work to look for the period of the day when the learner publishes a lot of publications in his social network account, we work in

our case with Facebook[1], so the period of activity is the period of the day when the learner publishes a lot in his Facebook account.

We define three periods of the day, from 8 am to noon, from 2 pm to 6 pm and from 6 pm to 10 pm. The period in which we find a large number of learner publications is the **Period of Activity (PA).**

$$PA = \max P_i \tag{5}$$

Where,

- i= 1, 2 ou 3
- P_i is the number of publications in the period number i (1=from 8 am to noon, 2=from 2 pm to 6 pm and 3=from 6 pm to 10 pm)

B. Sentiment Analysis

The second thing to do after the calculation of the period of activity is to recover all the publications that are published in this period and after filtering the publications according to the field of education, the idea is to recover only the Publications that contain terms related to education such as school, education, learn, learning, teacher, teaching, university, faculty, etc.

After we classify the publications obtained according to three classes: motivate, demotivate and neutral, using the dictionary AFINN[2] and after, the feeling of the learner in the period of activity takes the value of the majority class of publications for example if we have 20 publications that are related to education in the period of activity, and 10 of them are of motivated class, 6 of demotivate class and 4 of neutral class, so the majority class here is the motivate class and then the learner's feeling takes on the value motivate.

V. Description for Our Work

Our work as mentioned previously consists in proposing a new e-learning system by combining the use of genetic algorithms, the notions of information retrieval systems, social networks, sentiment analysis and Big Data with the Hadoop framework.

In this section, we will present the different stages of our work, how we combine genetic algorithms with the notions of information retrieval and apply the results to give the learner a pedagogical content corresponding to his profile, and how we use the Facebook's publications of the learner for adaptation, this work will be made using the MapReduce programming model and the HDFS (Hadoop Distributed file system).

[1]http://www.Facebook.com

[2]AFINN is a dictionary that contains words with weights between -5 and 5 which expresses the sentimental degree of the word

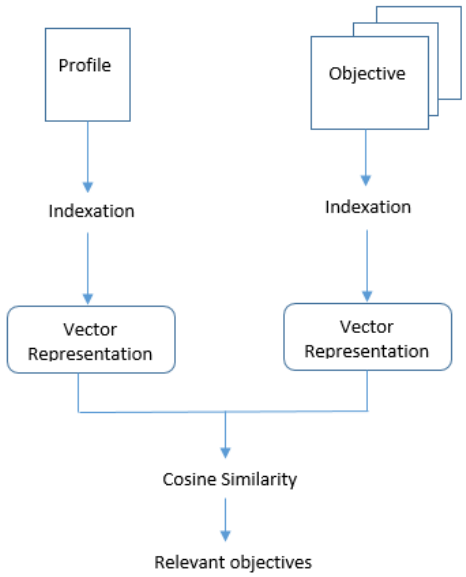

Fig. 1. Step of our information retrieval system.

A. Our Work: Information Retrieval

As we have said earlier the use of the field of information retrieval is important in our work either to evaluate the initial population for the genetic algorithm (GA) or for the reproduction of a new generation when applying the GA.

As our job is to find a pedagogical content for a learner profile the first thing to do is to save the profile as well the different pedagogical objectives for a course in text files, the idea is that we go to start looking for the pedagogical objectives those are relevant to the learner profile,it is like we want to find the relevant documents for a query expressed by a user, in our case the profile of the learner is like a query.

So at the beginning it is like an information retrieval system (IRS) we will index the learner's profile and the pedagogical objectives by creating representative vectors using the vector model [2] where each element is the weight of the term or concepts in the documents (profile or pedagogical objective), we will calculate the similarity between the learner profile and each pedagogical objective using Cosine similarity, Fig. 1 indicates the different steps to find the relevant objectives for the learner's profile.

B. Our Work: Genetic Algorithms

The second step in our work is the application of the genetic algorithm to find a single pedagogical objective that is optimal for the learner profile by pressing on the result of the information retrieval system to find the initial population.

After we calculate the similarity between the learner's profile and each pedagogical objective, we sort the result obtained in decreasing order to keep the first eight objectives that will play the role of the initial population this stage is called the selection stage of individuals who are most adapted to the working environment of the genetic algorithm.

For our problem of the research of the optimal pedagogical objective for the learner profile, individuals are the pedagogical

objectives and the fitness function to evaluate an individual is the cosine similarity.

We use to code the chromosome of an individual the binary coding, each chromosome's gene presents a concept of the course for example for the course of JAVA programming; classes is a concept, heritage is another concept, etc. Each gene takes either the value 1 if the concept exists in the pedagogical objective or 0 if it does not exist, with the size of each chromosome (number of genes) is equal to the number of concepts present in all the pedagogical objectives (our corpus). We find after the coding each pedagogical objective and also the profile of the learner with the equivalent chromosome where each gene (element of chromosome) takes either the value 0 or 1.

The chromosome is equivalent to the vector representation from a point of view of the information retrieval systems because after each generation of the GA we calculate the similarity between the profile and each individual (objective) of the generation.

The different steps to find the optimal pedagogical objective for the learner's profile are:

- **Choice of fitness function:**
  The choice of the fitness function is important in genetic algorithms, because it is the one that will give us an idea of an individual if it can be chosen or not, and also to find an optimal solution for the problem studied. In our work and since we want to work with the notions of information retrieval systems, we have chosen to use cosine similarity as a fitness function.
- **Selection of the initial population:**
  It is the step of computing the cosine similarity between the profile and each pedagogical objective after we sort the result in decreasing order to find the first 8 objectives which are the initial population where each pedagogical objective represents an individual using as criterion of evaluation (fitness function) the cosine similarity.
- **Coding of individuals:**
  The coding used in our work is the binary coding where each element of the chromosome (gene) takes either 0 or 1, 1 if a concept is present in a pedagogical objective and 0 otherwise, each pedagogical objective represents an individual of the population and each chromosome represents a vector where its size is equal to the number of concepts in all pedagogical objectives.
- **Crossover operation:**
  Another step which is also very important is the crossover which consists of mixing the characteristics of two parents of the previous generation to build two other children, we use in our work the crossover at a single point.
- **Mutation operation:**
  This operation consists of changing the value of a gene to increase the similarity value of its chromosome to be more adaptable to the working environment.

```
Generation 10:
==============
Population:
1 - 1110101000000011111111111110000000000000000000000000000000000000000100000000
2 - 1010000000000000000011000001000000010000000000000000000000000000000000111111111
3 - 1110101000000011111111111100000000000000000000000000000000000000000100000000
4 - 1110101000000011111111111110000000000000000000000000000000000000000100000000
5 - 1010000000000000000011000001000000010000000000000000000000000000000000111111111
6 - 1110101000000011111111111110000000000000000000000000000000000000000100000000
7 - 1110101000000011111111111110000000000000000000000000000000000000000100000000
8 - 1010000000000000000011000001000000010000000000000000000000000000000000111111111

Fitness:
1 - 0.2655565142631531
2 - 0.2949399517974853S
3 - 0.4680365025997162
4 - 0.15115077793598175
5 - 0.045195359736680984
6 - 0.12188316881656647
7 - 0.13612929866549377
8 - 0.11751154065132141

 meilleure solution de la generation 10:
 1110101000000011111111111110000000000000000000000000000000000000000100000000
```

Fig. 2. Result of our genetic algorithm.

The criterion of stopping our genetic algorithm is the number of generation fixed at the beginning, each generation is an iteration of the algorithm at each iteration we find the individuals of the population with new values of the similarity.

Our solution is a pedagogical objective (chromosome) exists in the last generation, it is the chromosome of the final population that has the greatest value of the similarity.

Fig. 2 shows an example of the result obtained by applying our approach to find the optimal pedagogical objective for a learner's profile:

From Fig. 2 it can be seen that our criterion for stopping the genetic algorithm is the number of iteration or generation which is equal in our case 10. As I said previously, our goal is to find the optimal objective in the last iteration, so according to Fig. 2, the program returns the optimal chromosome for the learner profile that has the high value of the fitness function (cosine similarity). And as an interpretation of the result obtained the learner must learn the concepts equivalent to the genes which have the value 1.

Fig. 3 gives the different steps of our genetic algorithms to find the optimal pedagogical objective.

C. Our Work: Social Networks and Sentiment Analysis

1) Period of Activity: As we have presented previously, the first thing to do after finding the optimal pedagogical objective for the learner profile is to calculate the period of the day where the learner must learn, for that we have proposed to use the facebook account of the learner or the twitter account to know the period when the learner publishes a lot, it is therefore at this period when the learner is very active and therefore it is at this period that the learner can Improve his/her level of learning easily.

To calculate the period of activity we use an API called RestFB[3] Which allows us to retrieve all the information about the learner who connects to the platform using his Facebook account, like his name, first name, email, and his publications, etc. And the most important thing is that the RestFB API Gives us the possibility to recover at what time the publication is published on Facebook which allows us to easily find the period of activity of the learner.

[3]RestFB (http://restfb.com/) is a simple and flexible Facebook Graph API client written in Java. It is open source software released under the terms of the MIT License.

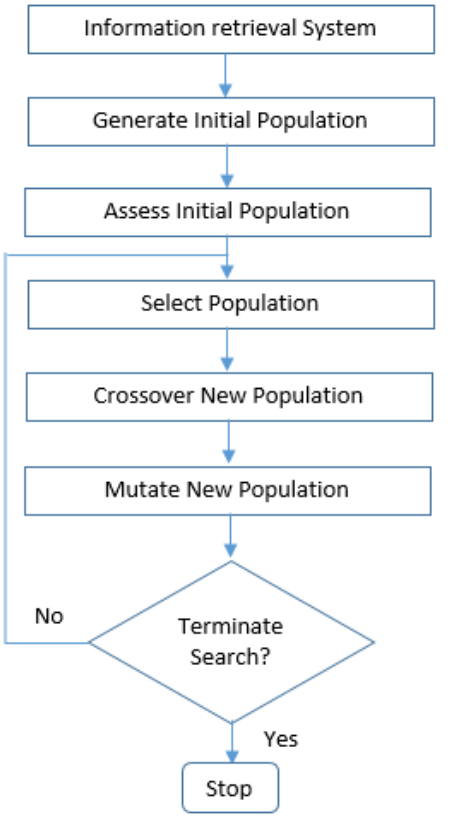

Fig. 3. The process of our genetic algorithm.

Fig. 4. Facebook application.

The RestFB API needs some configuration like creating a Facebook application in the Facebook developers space [4] this application contains a name, a working domain name must be the same domain name for our platform, an application ID and apps secret ID, the last two parameters are the ones that give us the ability to collect data from the learner account, Fig. 4 illustrates a Facebook application under the name madani:

The second step which is very necessary for using the RestFB API is to install the login button that is to say, make a connection between the facebook application using one of the SDK proposed by Facebook like PHP SDK or JavaScript SDK, in this work we tried to use JavaScript SDK with JavaEE.

After creating the Facebook application and the login button we can retrieve a parameter that is used by RestFB and which allows us to retrieve the learner's publications for using it in the search phase of the period of activity, it is the ACCESS TOKEN.

[4]https://developers.facebook.com/

After all these steps we will have the publications of the learner, and by using the API RestFB and the formula 5 we can easily find the period of activity.

2) Sentiment Analysis: After we find the period of activity of the learner (the period of the day when the learner was very active), we collect all the publications that are published in this period and filter them in relation to publications that are related to education, to do this work we use a list of terms related to education such as learning, learn, teacher, school, university, students, faculty, education, teaching, training, etc.

The idea is that among the publications of the period of activity we keep only those are in relation to teaching and education, and afterwards we classify these publications according to three classes: motivate, demotivate or neutral.

For the classification, we use the AFINN dictionary, but before the classification process we have to make text pre-processing on the publications to decrease the noise that will influence the classification, such as:

- **Tokenization:** Which is the phase of splitting the tweet into terms or tokens by removing white spaces, commas and other symbols etc, it's an important step because in our works we focus on individual words to search them is the AFINN dictionary or in WordNet.
- **Removing Stop Word:** It removes The words that have no effect on the classification of tweets like preposition and the article (a, an, the).
- **Removing URL:** The URLs have no effect on the classification so it is important to eliminate them.
- **Removing numbers**
- **Stemming:** Stemming is another very important process. In our work and because we focus on English language we use the porter stemming [19].

After the text pre-processing phase, For a given publication of the learner in its Facebook account we go through its words and we look for their weights from the AFINN dictionary, after we sum the weights of all words of the publication those exist in AFINN, after that we calculate the average of the weights and using a threshold we classify the tweet according to three classes Motivate, demotivate or neutral.

Our proposal is to work with dictionary-based approach, for that, we use the AFINN dictionary which contains words in English with a weight that can take a value between 5 and -5 (strongly positive, mildly, strongly negative, etc.) for example the word "abandoned" has the value -2 and the word "accept" has the value 1, etc. and to look for the feeling of a publication the idea is to browse all the words in it and sum these weights using AFINN and after using a threshold to classify the publication into motivate, demotivate or neutral.

In Table I, we present the classification and error rate with the use of AFINN before and after the application of text pre-processing (TP), and from this table, we remark that the text pre-processing increase the classification rate and decrease the error rate.

TABLE I. CLASSIFICATION AND ERROR RATE

Method	Classification rate	Error rate
Without TP	0.57	0.43
With TP	0.76	0.24

VI. PARALLELIZATION OF OUR WORK

In an e-learning platform it is possible to find many learners also many courses and for each course many chapters and concepts, so a large amount of data which makes learning in the platform ineffective because of the time that we can wait to find the result of analyzing a learner's profile, also today everybody publishes in social networks so for a learner it is possible that we find a lot of publication in his facebook account (huge volume of publications) so the analysis of his publications and his profile becomes difficult with the classic means.

Our proposal which is the main work of this article is to parallelize our genetic algorithm to find the optimal pedagogical objective for the profile of the learner, also we parallelize the classification of the publications for finding the sentiment of the learner in the period of activity.

We propose to work in a big data system using HADOOP's solutions; Hadoop Distributed File System (HDFS) and the programming model MapReduce Which is a Java API for writing distributed programs for information retrieval ,The idea is to install a cluster Hadoop (set of machines) and to share the work of the analysis of the learner's publications and the search for the optimal pedagogical objective between the machines of the cluster.

A. Search for the Optimal Pedagogical Objective

1) Description: For that we record the profile of the learner as well as the pedagogical objectives for a course in text files and save them in HDFS to parallelize the storage between the different machines of the Hadoop cluster, also we record the result of the analysis of the learner's profile in HDFS as the optimal chromosome.

To optimize the storage volume of the data in HDFS we stored all the pedagogical objectives in the same file instead of giving each objective a file because the file system (HDFS) will give for each file the size of the block which is 64MB therefore for example if a file has a size of 10MB after storage in HDFS its size will be 64MB therefore an increase in the size of the stored data, so we put all the objectives in the same file to optimize the storage size.

After a learner passes a QUIZ for such course, he gives us an idea of the concepts that he does not master and it is these concepts that will form his profile, after we save these concepts in a text file, the same applies to each pedagogical objective we save it in a file with the name and email of the teacher that he puts it on the platform, so in the document of the pedagogical objectives we find each objective with a name of teacher, his email And the concepts that the learner must learn to validate the course.

The different parallelization's steps of our work are the following:

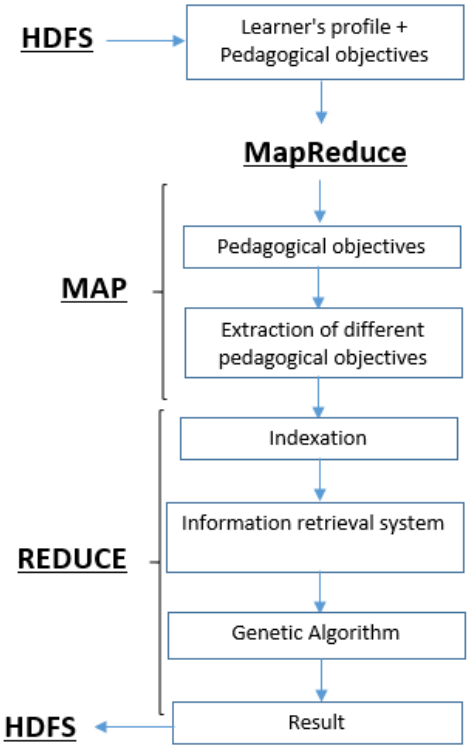

Fig. 5. Steps of parallelization.

- **The recording of the data to be analyzed:**
  As I have already said we record the learner profile and the pedagogical objectives (same text files) in HDFS.
- **MapReduce program:**
  The parallelization of our work is done with the programming model MapReduce which is a Java API allowing the writing of the distributed programs, it Constitutes of two main operations, the Map operation and the Reduce operation.
  In our case, the Map method allows us to browse the file containing the learning objectives and to extract all the pedagogical objectives after sending the result to the Reduce method which allows us to index the learner's profile and Pedagogical objectives, the application of the information retrieval system and the application of the genetic algorithm in a parallel manner.
- **Recording of the analysis result:**
  The last step is the recording of the result obtained after the MapReduce operation in HDFS which is the optimal pedagogical objective for the learner's profile in the form of a chromosome formed by genes with either the value 1 meaning that The learner must learn the equivalent concept, or the value 0 otherwise.

2) Schematization of Parallelization Steps: Fig. 5 gives the different steps for parallelizing the adaptive learning algorithm.

3) MapReduce Algorithm: Our MapReduce algorithm followed To find the optimal goal for a learner is the following:

Class Mapper

Method Map(Docid,FIleOfObjective)
For each line ∈ FIleOfObjective
Write(Docid,line)
End for

Class Reducer

Method Reduce(Docid,List(line))
$S \leftarrow NULL$
For each n ∈ List(line)
$S \leftarrow S + n$
End for
$List2 \leftarrow Split(s)$
Index(List2 & Learner's profile)
$Result \leftarrow BuildingGA()$
Write(Result," ")

With:

- **Split(s):** Extracts the various pedagogical objectives.
- **Index(List2 & Learner's profile):** A function which makes the vector representation of each pedagogical objective and of the learner's profile.
- **BuildingGA():** The execution of the genetic algorithm

B. Search for Sentiment of the Learner

After we collect the Facebook publications of the period of activity either with the RestFB API or with Apache Flume we store them directly in HDFS to parallelize the classification.

1) Parallelization Steps: Fig. 6 shows the different steps to parallelize the classification using HDFS and MapReduce.

2) Description of Parallelization: From Fig. 6 the first step of the parallelisation is to store the data set (publications of the learner) to classify in HDFS for sharing the storage between several machines (Hadoop cluster) and after it is the step of classification by applying The AFINN method.

The input of the MapReduce operation at each iteration contains a publication to classify and the output contains the sentiment of the learner, the result of the classification is stored in HDFS.

Our MapReduce algorithm followed for the classification of the publications is the following:

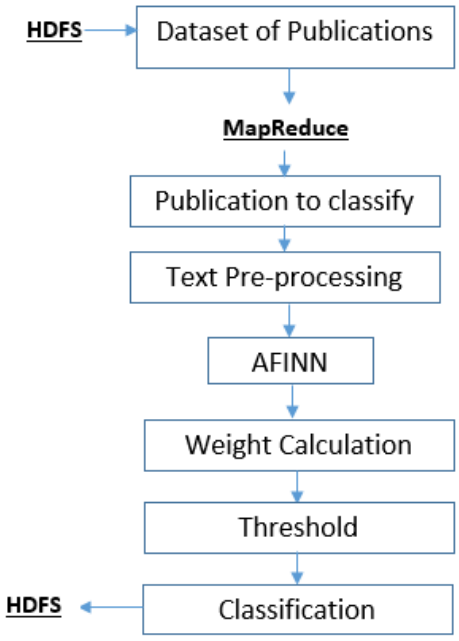

Fig. 6. Parallelization steps.

Class Mapper

Method Map(publication)
S ← 0
publication ← TextPreProcesssing(publication)
SE[] ← Split(publication)
For each word ∈ SE
S ← S+AFINN(word)
End for
sentiment ← s/(length(SE))
write(sentiment,1)

Class Reducer

Method Reduce(sentiment,ListOfOne(1,1,1...))
S ← 0
For each n ∈ ListOfOne(1,1,1...)
S ← S+n
write(sentiment,S)

Where,

- TextPreProcesssing(publication): Apply different types of text pre-processing
- AFINN(word): Calculate the weight of feelings equivalent to word if it is existent in AFINN

After the calculation of the period of activity (PA) which will give us an idea of the part of the day when the learner must learn, the classification of the learner's publications which are published in PA according to three classes (motivate, demotivate, neutral) gives the platform's teacher an idea of the level with which the learner must learn and it depends on

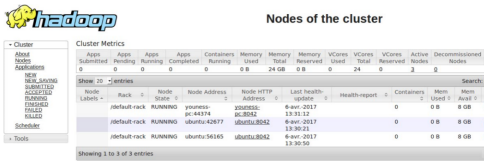

Fig. 7. Configuration of our cluster.

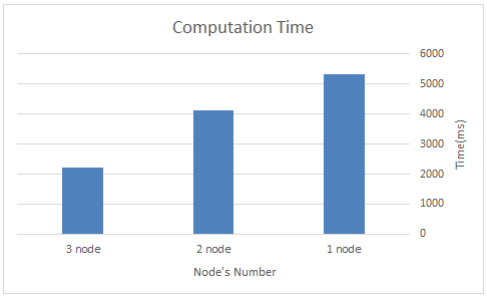

Fig. 8. Calculation time.

his motivation, of course, if he is motivated we can increase the level of study, and the opposite if he is demotivated.

VII. Parallelism Effect

As we said earlier our work consists of parallelizing the search for an optimal pedagogical content for the learner in a big data system using the Hadoop framework with Hadoop Distributed File System (HDFS) to distribute the data storage necessary for the Analysis (pedagogical objectives, learner profile and the learner publications), and to record the result of the analysis (the optimal chromosome and the sentiment of the learner) to facilitate the interpretation afterwards and also to parallelize the classification of the publication.

The goal of parallelizing the work is to reduce the computation time if we have a large size of the data (profile, objectives and publications), for this and to demonstrate the effect of sharing the work between several machines we decide to work with a variable number of nodes (Hadoop machines).

A. Configuration of our Cluster

Fig. 7 shows the configuration of our cluster (Hadoop machine set) which contains three Hadoop machines, one master machine and two slave machines, so our work will be shared between three machines to reduce the computation time. The nodes used are three nodes with the Linux UBUNTU 15.04 operating system.

According to Fig. 7, our Hadoop cluster contains three UBUNTU machines that are all three enabled with a total storage of 24GB (8GB for each node).

B. Effect of Parallelization

Fig. 8 represents the calculation time of our work using respectively 1, 2 and 3 Hadoop nodes.

From Fig. 8 we notice that by increasing the number of nodes in the cluster the computation time decreases and this is the goal of sharing the work between several machines.

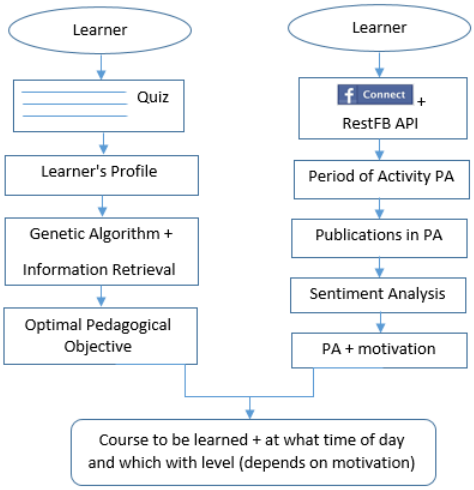

Fig. 9. Different steps of our adaptive e-learning system.

A very important information to note is that a MapReduce program which works with one node works also with three nodes and can also work with a cluster of 100 nodes, so if we want to decrease the calculation time just we need to add another node to the cluster which is a very easy operation.

VIII. Summary of our Adaptive E-learning System

Fig. 9 presents the different steps to adapt a course to a learner based on his/her profile, the period of activity and his motivation.

According to Fig. 9, our system is based on two levels of adaptation, the first one is in the form of an optimization problem which makes it possible to find the pedagogical objectives which are optimal in relation to the learner's profile, and the second is to use the data of the learner published in his social networks accounts to find his period of activity and his motivation related to education (publications that are related to education).

At the end of all these processes, we find the courses that the learner must learn, at what period of the day (the period when the learner is active) and with what level (depends on his motivation).

IX. Conclusion and Perspectives

In this paper, we have presented our system of adaptive learning based on a genetic algorithm, the notions of information retrieval systems, Data from social networks,New measure called period of activity, sentiment analysis and Big Data technologies (Hadoop, HDFS, MapReduce); by searching the pedagogical content that is optimal for a Learner based on his profile and a set of pedagogical objectives set by a teacher,the period of activity (when the learner is active) and with what level the learner must learn (his motivation).

In this work, we have proposed a measure called the period of activity that it helps us to find the period of the day when the learner is active (publish a lot of publications in his social networks accounts).

We proposed also a new approach to parallelize the algorithm which allows to search the optimal pedagogical objective and for finding the motivation of the user by using a Hadoop cluster of three-node (UBUNTU machines) to share the work between several machines.

With the use of a Hadoop cluster the computation time decreases by increasing the number of nodes in the cluster which facilitates the search of the pedagogical content for the learner if we have a large volume of the data to be analyzed,and also facilitate the search of the motivation of the learner if we have a lot of publications in his social network account.

Our next work is to develop a new approach for the sentiment analysis without the use of AFINN dictionary based on the semantic, and also use a new optimization algorithm of artificial intelligence to find a course adapted to the learner's profile.

REFERENCES

[1] Samia Azough, Mostafa Bellafkih and El Houssine Bouyakhf,Adaptive E-learning using Genetic Algorithms,IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.7, July 2010

[2] Salton, G. Automatic text processing: the transformation, analysis, and retrieval of information by computer. New York: Addison- Wesley Publishing Co. Inc. 1989.

[3] David, L. Handbook of Genetic Algorithms. New York : Van Nostrand Reinhold. 1991.

[4] BANGORN KLABBANKOH,OUEN PINNGERN PH.D.,APPLIED GENETIC ALGORITHMS IN INFORMATION RETRIEVAL Md. Abu Kausar,Md. Nasar and Sanjeev Kumar Singh,A Detailed Study on Information Retrieval using Genetic Algorithm,Journal of Industrial and Intelligent Information Vol. 1, No. 3, September 2013,doi: 10.12720/jiii.1.3.122-127

[5] Laith Mohammad,Qasim Abualigah and Essam S. Hanandeh,APPLYING GENETIC ALGORITHMS TO INFORMATION RETRIEVAL USING VECTOR SPACE MODEL,International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.5, No.1, February 2015

[6] Vrajitoru, 1998,Crossover improvement for the genetic algorithm in information retrieval,Information Processing and Management: an International Journal Volume 34 Issue 4, July 1, 1998

[7] Philomina Simon and S. Siva Sathya,Genetic Algorithm for Information Retrieval,International Conference on Intelligent Agent & Multi-Agent Systems,IAMA 2009, doi: 10.1109/IAMA.2009.5228033

[8] Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, and Osman A. Sadek,Using Genetic Algorithm to Improve Information Retrieval Systems,International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:2, No:5, 2008

[9] W. Fan, M.D. Gordon and P. Pathak. Personalization of search engine services for effective retrieval and knowledge management, in: Proc. 2000 International Conference on Information Systems (ICIS), Brisbane, Australia, 2000.

[10] Hawkes, L. W., Derry, S. J. Advances in local student modeling using informal fuzzy reasoning. International journal of human-computer studies, 45(6), 697-722,1996

[11] Ruiz, M. D. P. P., Barriales, S. O., Prez, J. R. P., Rodrguez, M. G. Feijoo. net: an approach to personalized E-learning using learning styles. In Web Engineering (pp. 112-115). Springer Berlin Heidelberg,2003

[12] Chrysafiadi, K., Virvou, M. Evaluating the integration of fuzzy logic into the student model of a web-based learning environment. Expert Systems with Applications, 39(18), 13127-13134,2012

[13] Martin, J., VanLehn, K. Student assessment using Bayesian nets. International Journal of Human-Computer Studies, 42(6), 575-591,1995

[14] Vicari, R., Flores, C. D., Seixas, L., Gluz, J. C., and Coelho, H. AMPLIA: A Probabilistic Learning Environment. International Journal of Artificial Intelligence in Education, 18(4), 347-373,2008

[15] Romero, C., Ventura, S., and De Bra, P. Knowledge discovery with genetic programming for providing feedback to courseware authors. User Modeling and User-Adapted Interaction, 14(5), 425-464,2004

[16] Chang, T. Y., and Ke, Y. R. A personalized e-course composition based on a genetic algorithm with forcing legality in an adaptive learning system. Journal of Network and Computer Applications, 36(1), 533-542,2013

[17] Manju Bhaskar, Minu M Das,Dr. T. Chithralekha and Dr. S. Sivasatya,Genetic Algorithm Based Adaptive Learning Scheme Generation For Context Aware E-Learning,Manju Bhaskar et. al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010, 1271-1279

[18] D. Ganesh & V. V. RamaPrasad, Protection of Shared Data Among Multiple Users for Online Social Networks,International Conference on Contemporary Computing and Informatics (IC3I), 2014.

[19] M.F.Porter,An algorithm for suffix stripping,Originally published in Program, 14 no. 3, pp 130-137, July 1980.