# Towards an Approach Based on Adjusted Genetic Algorithms to Improve the Quantity of Existing Data in the Context of Social Learning

https://doi.org/10.3991/ijet.v16i09.20685

Sonia Souabi (✉), Asmaâ Retbi, Mohammed Khalidi Idrissi,
Samir Bennani
Mohammed V University, Rabat Morocco
souabisonia@gmail.com

**Abstract**—In the current era, multiple disciplines struggle with the scarcity of data, particularly in the area of e-learning and social learning. In order to test their approaches and their recommendation systems, researchers need to ensure the availability of large databases. Nevertheless, it is sometimes challenging to find-out large scale databases, particularly in terms of education and e-learning. In this article, we outline a potential solution to this challenge intended to improve the quantity of an existing database. In this respect, we suggest genetic algorithms with some adjustments to enhance the size of an initial database as long as the generated data owns the same features and properties of the initial database. In this case, testing machine learning and recommendation system approaches will be more practical and relevant. The test is carried out on two databases to prove the efficiency of genetic algorithms and to compare the structure of the initial databases with the generated databases. The result reveals that genetic algorithms can achieve a high performance to improve the quantity of existing data and to solve the problem of data scarcity.

**Keywords**—Social Learning, Data Scarcity, Genetic Algorithms

## 1 Introduction

E-learning is proving to occupy a major position in providing online courses and education [1]. As learning practices evolve, social learning allows e-learning to foster collaboration and interactivity among online learners. Currently, the concept of social learning is generally associated with social networks, which are experiencing a high level of activity as a result of high user demand [2, 3, 4]. In this respect, we recognize the relevance of implementing a recommendation system within a social learning network. Indeed, recommendation systems enable the framing of a learner's work and offer relevant material according to the profile, the interactivity, the popularity of the content, and many other aspects that can be incorporated in a recommendation system [5]. Collaborative filtering and hybrid approaches are among the recommendation system techniques considered by researchers in terms of e-leaning and social learning [6, 7]. Yet e-learning environment suffers from the scarcity of data. In order to properly test recommendation approaches, access to a large database is required. Indeed, the performance of Machine Learning algorithms decreases when the database is not of considerable volume. In this regard, we propose genetic algorithms with some adjustments as a solution to increase our data and to test our recommendation approach in a more trustworthy way. We have further envisioned a method to increase our data beyond the synthesis since genetic algorithms are based on creating new chromosomes, and thus creating new individuals and obtaining diversified data. We have thus emphasized the diversification feature which is considered among the advantages of genetic algorithms. We perform the test on two databases of different sizes, and afterwards we apply genetic algorithms to generate additional data. For comparison, we evaluate the box plots owned by the generated databases. The intention is to compare the structure of the initial databases with the generated databases.

The paper is spread out as follows. The first part consists of an overview of genetic algorithms in the e-learning field. The second part includes the methods and tools involved in our study. The third part focuses on the results obtained through the tests performed. The following section presents a discussion of the results achieved. Finally, a conclusion summarizes all the work performed and the next directions to be pursued.

## 2 Background

### 2.1 Related works

In recent years, researchers have increasingly turned their attention to exploring the Learning Machine within e-learning. Indeed, algorithms can indeed be used to solve a number of learning difficulties. Among the algorithms that have been explored in e-learning: genetic algorithms. Ahmed hamdi Abu absa, Sana’a wafa al-sayegh (2008) deal with an important problem, which is the scheduling of classes [8]. Genetic algorithms are thus considered in order to solve the problem of presentation of the course timetable. NEBOJŠA GAVRILOVIĆ, TATJANA ŠIBALIJA (2018) provide a state of the art on the use of evolutionary algorithms in distance learning [9]. The study also shows the relevance of genetic algorithms in the personalization of the learning path. V. V. Zaporozhko and I. P. Bolodurina (2018) propose the evaluation of genetic algorithms in the construction of a learning path that is optimal and beneficial for learners [10]. Akure, Ondo State, Nige-ria, O. C. Agbonifo, and O. A. Obolo (2018) allow the usage of genetic algorithms to generate an optimal path through the identification of the level of difficulty of online courses and the courses that correspond to the learners' needs [11]. Queiroga, E.M.; Lopes, J.L.; Kappel, K.; Aguiar, M.; Araújo, R.M.; Munoz, R.; Villarro-el, R.; Cechinel, C (2020) propose to exploit genetic algorithms to predict at-risk students in a Brazilian university [12]. The solution developed includes the integration of the different interactions of learners within the virtual learning environment. Amelec Viloria, Mercedes Gaitan Angulo, Sadhana J. KamatkarJuan de la Hoz – Hernandez, Jesús García Guiliany Amelec Viloria, Osman Redondo Bilbao, Hugo Hernandez-P (2020) seek to study learners' interactions with learning materials in order to assess the quality of the material as well as the learning process [13]. The method used, therefore, proposes to compare genetic algorithms with association rules and the decision tree.

Several studies merely proposed intelligent systems in the e-learning field regardless of the problem of data scarcity [14]. Many researchers restrict their efforts to testing approaches on small databases of non-significant size, which raises a genuine problem for machine learning algorithms as long as they require large databases. Many recommendation systems were tested only on databases with no more than 100 learners [15, 16, 17], whereas we require a lot of other data to measure the performance of a recommendation system. In what follows, we intend to combine genetic algorithms with our recommendation system in order to solve the problem of data scarcity. We intend to orient our work towards another path different from the previous research contexts, namely genetic algorithms to solve the data scarcity.

### 2.2 Genetic algorithms background

The genetic algorithm starts with a set of solutions (denoted by chromosomes) called population. The selected solutions form new solutions based on their fitness value - the more the value of fitness, the more chances they have to reproduce [18]. The Basic Genetic Algorithm has been explained in Algorithm 1.

**Algorithm 1 : Basic genetic algorithm**

```
1: Choosing the initial population presented by chromosomes
2: Estimating the fitness of every chromosome
3: Generating the new population
Selection: Choosing chromosomes according to the fitness function.
Crossover: Applying crossover to every chromosome parent.
Mutation : Applying mutation to every chromosome parent generated by crossover
4: If the terminating condition is satisfied, end the process and return the result.
```

Many studies have addressed genetic algorithms in the context of education and learning [19, 20]. We also address genetic algorithms in the context of social e-learning, however, in order to increase the quantity of data coming from an available database. In this way, we will face the problem of data scarcity in order to test the proposed approaches in distance learning.

## 3 The Proposed Approach

In the context of social learning, the scarcity of data raises a serious challenge in testing different approaches, including recommendation systems. In order to carry out reliable and effective testing, a wide and sufficient starting base is necessary. To address this problem, increasing data is a radical solution by generating additional data from existing data. To generate additional data, however, the data must meet certain criteria to ensure that it behaves in the same way as the original data. In this respect, we propose to apply genetic algorithms to generate additional data and then compare the generated database with the initial database. We will therefore compare the characteristics of the two databases in order to find out to how far genetic algorithms were able to generate data similar to the initial database. The algorithm below (algorithm 2) highlights the modified version of the genetic algorithms considered in our proposal.

**Algorithm 2 : Adjusted version of genetic algorithms**

```
1: Choosing the initial population as selected chromosomes
Constituting the possible parent chromosomes according to the original population.
Crossover: Applying crossover to every chromosome parent.
Mutation: Applying mutation to every chromosome parent generated by crossover.
2: End of the process and return the new population generated from the initial population.
```

Genetic algorithms are generally applicable in an optimization context, but our application context is not part of an optimization problem, but mostly a problem of data scarcity. Since genetic algorithms allow us to generate new individuals from existing individuals, we thought of applying them to generate new data.

To improve the quantity of the existing database, we propose in our approach to apply genetic algorithms, though with some adjustments in order to adapt them to our context to solve the problem of lack of data. The figure below shows the process pursued (fig. 1).

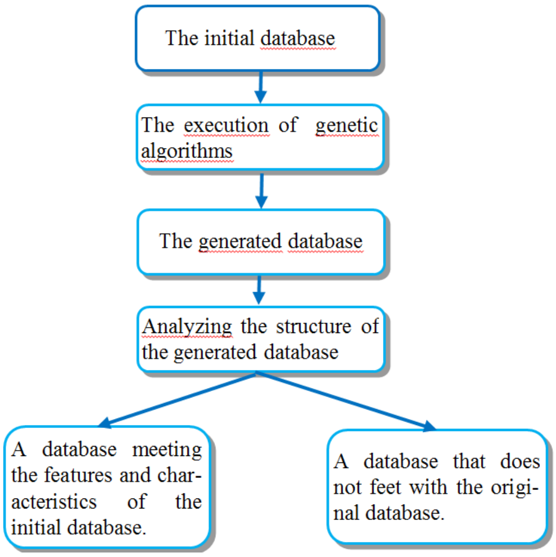
The initial database
The execution of genetic algorithms
The generated database
Analyzing the structure of the generated database
A database meeting the features and characteristics of the initial database.
A database that does not feet with the original database.

**Fig. 1.** The process of genetic algorithms model for improving the quantity of an existing data

## 4 Tests and Results

In order to analyze the results obtained in both cases on the basis of genetic algorithms, it is initially required to define the different parameters under consideration. Both crossing and mutation operators are implemented with a probability of 1, and the number of iterations is limited in a single iteration (n=1). By choosing the number of iterations equal to 1, a database of size 30 is generated from the first database, and a database of size 90 is generated from the second database. To reveal the efficiency of genetic algorithms, a box plot is drawn to compare the initial databases with the generated databases. The box plot actually allows to show the essential profile of a statistical series containing specific data. It summarizes some position indicators:

- Median
- Quartile
- Minimum
- Maximum

The box plot enables to draw up the statistical characteristics of a database. By comparing the box plot from the initial database to the generated one, we assess the capacity of genetic algorithms to generate behavioral data similar to the initial database.

We perform the test on two databases within the social learning context:

- The first database: In a social learning group on Facebook, we collect the number of reactions and comments per post related to Mathematics for a period of one month in order to evaluate the interactivity rate of learners within a social learning environment. Leaners age is between 17 and 18 years old.
- The second database: In the same social learning group on Facebook, we collect the number of reactions and comments per post related to Physics for a period of one month. The purpose is to evaluate the interactivity of learners within Physics.

To test genetic algorithms, we will use the R programming language while applying the different operators of genetic algorithms.

### 4.1 First database

A group of learners between the age of 17 and 18 interact in a Facebook social learning group on several content areas, including math (Table 1). For a period of one month, data is collected by examining the number of reactions and the number of comments per post. Informations about the first database are as follows:

- Two characteristics (number of reactions per post and number of comments per post).
- A database of size 6 (informations about 6 posts).

**Table 1.** First initial database

| Post | Number of reactions | Number of comments |
|---|---|---|
| Post 1 | 80 | 2 |
| Post 2 | 40 | 3 |
| Post 3 | 100 | 1 |
| Post 4 | 70 | 4 |
| Post 5 | 22 | 5 |
| Post 6 | 60 | 3 |

By running genetic algorithms, we obtain a database with the following properties while considering the following parameters:

$$P_c = 1$$
$$P_m = 1$$
$$N_i = 1$$

$P_c$: Probability of crossover operator
$P_m$: Probability of mutation operator
$N_i$ : Number of iterations

As a result, we generate a database with the following informations (Table 2):

- A database with two characteristics (as the initial database).
- A database of size 30 obtained from the initial database (6×(6-1)=30).

**Table 2.** Additional data generated by genetic algorithms from the first database

| Number of reactions | Number of comments |
|---|---|
| 89 | 2 |
| 33 | 3 |
| 85 | 0 |
| 97 | 3 |
| 87 | 1 |
| 65 | 7 |
| 87 | 0 |
| 17 | 7 |
| 93 | 2 |
| 49 | 3 |
| 37 | 0 |
| 105 | 2 |
| 39 | 1 |
| 73 | 6 |
| 39 | 0 |
| 25 | 6 |
| 45 | 2 |
| 57 | 2 |
| 103 | 1 |
| 69 | 4 |
| 103 | 0 |
| 21 | 4 |
| 109 | 2 |
| 53 | 0 |
| 71 | 4 |
| 23 | 5 |
| 77 | 6 |
| 55 | 1 |
| 29 | 6 |
| 55 | 0 |

To evaluate the results, a box plot study is carried out on the initial database and the generated database for each type of data, namely the number of reactions and the number of comments (Fig. 2 and Fig. 3).

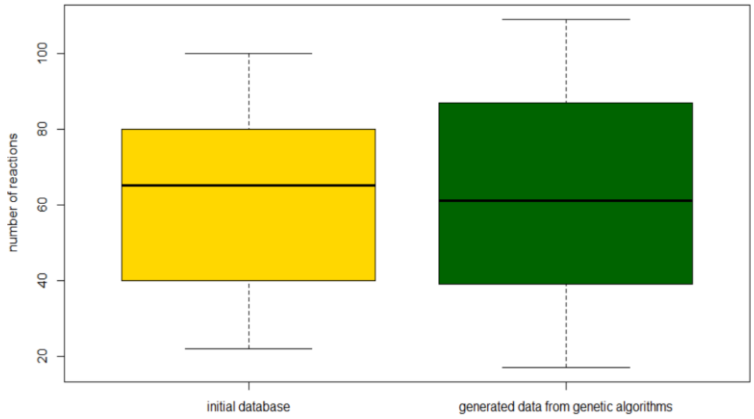

**Fig. 2.** The number of reactions according to the first initial database and the generated data from genetic algorithms

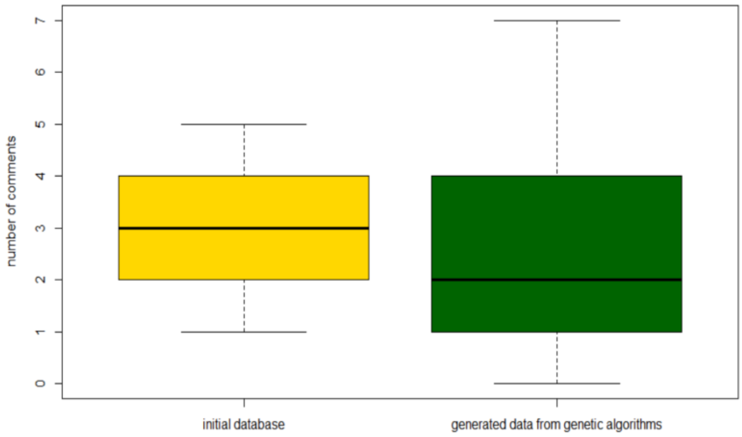

**Fig. 3.** The number of comments according to the first initial database and the generated data from genetic algorithms

A boxplot describes the statistical characteristics of a database and the dispersion of the data with respect to the median. In our case, we realize that the median of both figures remains approximately similar. As for the dispersion of values, genetic algorithms generate more dispersed data. For instance, the second figure shows a dispersion between 0 and 7 for the genetic algorithms versus a dispersion between 1 and 5 for the initial database. This shows that the application of genetic algorithms to improve the quantity of a database is highly recommended as they allow to keep approximately the same structure as the initial database.

### 4.2 Second database

Informations of the second database are provided as follows (Table 3):

- A database with two characteristics (number of reactions per post and number of comments per post).
- A database of size 10 (informations about 10 posts).

**Table 3.** The second initial database

| Post | Number of reactions | Number of comments |
|---|---|---|
| Post 1 | 6 | 4 |
| Post 2 | 3 | 4 |
| Post 3 | 6 | 3 |
| Post 4 | 5 | 5 |
| Post 5 | 3 | 6 |
| Post 6 | 4 | 3 |
| Post 7 | 2 | 0 |
| Post 8 | 1 | 1 |
| Post 9 | 6 | 1 |
| Post 10 | 5 | 4 |

By applying the genetic algorithms, we reach a database with the following characteristics while considering the following parameters:

$$P_c = 1$$
$$P_m = 1$$
$$N_i = 1$$

$P_c$: Probability of crossover operator
$P_m$: Probability of mutation operator
$N_i$ : Number of iterations
As a result, we generate a database with the following informations:

- A database with two characteristics (as the initial database).
- A database of size 90 obtained from the initial database (10×(10-1)=90).

Here is an excerpt of the results obtained by applying genetic algorithms to improve the quantity of the second database (Table 4). It was not possible to insert all the results obtained due to the size of the generated database (100 data).

**Table 4.** Part oft he results obtained after application oft he genetic algorithms on the second database

| Number of reactions | Number of posts |
|---|---|
| 3 | 5 |
| 7 | 3 |
| 5 | 4 |
| 5 | 7 |
| 5 | 3 |
| 3 | 1 |
| 1 | 1 |
| 7 | 1 |
| 5 | 5 |
| 6 | 3 |
| 6 | 4 |
| 2 | 7 |
| 4 | 3 |
| 3 | 1 |
| 0 | 1 |
| 6 | 1 |
| 6 | 5 |
| 5 | 6 |
| 3 | 6 |
| 5 | 2 |

From the moustache boxes obtained, it appears that the medians have approximately similar values. The only difference again lies in the dispersion of the values for the two figures. In fig. 4, 50% of data are between 2 and 6. For fig. 5, the 50% dispersion of the data is approximately similar with a minor difference in the maximum value.

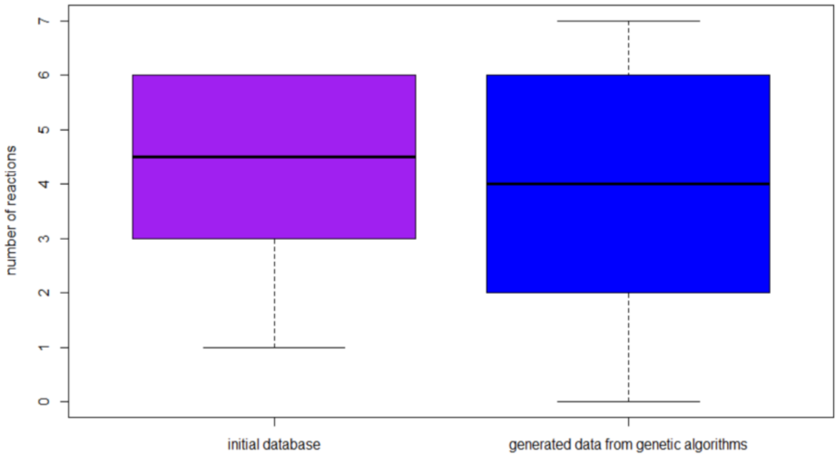

**Fig. 4.** The number of reactions according to the second initial database and the generated data from genetic algorithms

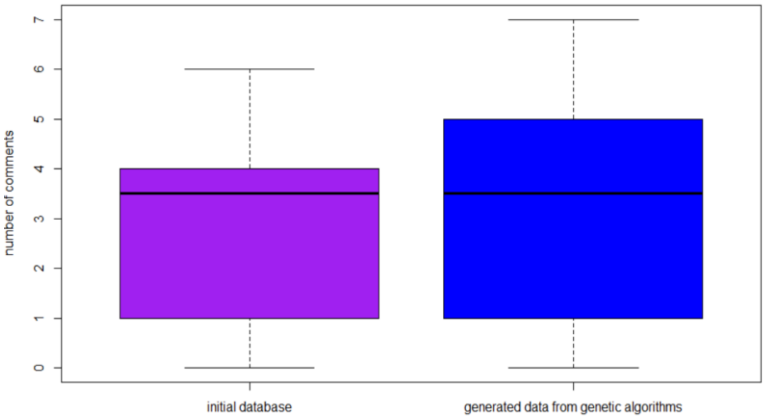

**Fig. 5.** The number of comments according to the second initial database and the generated data from genetic algorithms

Based on the databases on which we conducted our study, it appears that genetic algorithms provide results that match the behaviour of the real data. In addition, no significant change is noticed regarding the median. The median remains approximately the same between the initial database and the database generated by the genetic algorithms. Likewise for the max and min values, the difference remains minimal with a deviation of ±1 between the initial data and the simulated data. Another noteworthy point to report is that 50% of the data generated by genetic algorithms have a wider dispersion than the initial data.

Our adjusted version of genetic algorithms brings several advantages:

- Expanding our initial database.
- Create new diversified individuals from the initial individuals.
- Adapt the number of data generated according to the choice and the approach to be tested.

The reuse of our adjusted version of genetic algorithms can as well be adapted to other disciplines and areas affected by the problem of data scarcity. New individuals can be created from the initial population with the same behavior of the initial data.

In order to test a recommendation system within a social learning environment, it is important to initially set up a suitable database allowing the performance of the recommendation system to be assessed. In some cases, it is challenging to ensure a large and reliable database. Learner data is very limited and sometimes the number of learners is limited as well. So we need more data to enrich the process and practice the algorithms in a more concrete way. Consequently, we have to opt for an expanded database, i.e. improving the quantity of an existing database. From the research undertaken, it appears that genetic algorithms offer an appropriate solution for this problem. The data generated by the genetic algorithms match roughly the similar structure and features of the initial database. This is the purpose of our proposal, to offer a solution for the data scarcity provided as long as the generated data can keep the same properties of the initial database. This will give us the opportunity to test our recommendation approach while avoiding the scarcity of data.

So, from initial population, we can generate n×(n-1) new individuals. Database size increases significantly if we apply genetic algorithms on the initial database. For instance, an initial database of 100 individuals will generate 9000 new data, which means that genetic algorithms act as a very important part of data generation to test a certain approach.

## 5 Conclusion

In many circumstances, we are confronted with a very frequent problematic, namely the data scarcity. In order to leverage machine learning in the various areas, particularly e-learning, it is advisable to ensure a large database for effective testing of the proposed approaches. In this regard, the testing database ought to be sufficiently large to conduct reliable tests. This article proposes a solution based on genetic algorithms in order to improve the quantity of an existing database. We perform the test on two databases by comparing the structure of the initial database with the structure of the database generated by genetic algorithms. It turns out that genetic algorithms yield performing results in terms of statistical structure, notably for the median and the box plot. Our future work will consist of:

- Testing genetic algorithms to create larger databases.
- Exploit genetic algorithms within our recommendation approach to solve the data scarcity problem.

## 6 References

[1] Al-Fraihat, D., Joy, M., Masa'deh, R., Sinclair, J.: Evaluating E-learning systems success: An empirical study. Computers in Human Behavior 102, 67–86, (2020). https://doi.org/10.1016/j.chb.2019.08.004

[2] W. Elsayed: Students and the Risk of Virtual Relationships in Social Media: Improving Learning Environments. Int. J. Emerg. Technol. Learn., vol. 15, no. 21, p. 118, (2020). https://doi.org/10.3991/ijet.v15i21.15063

[3] Y. M. Al-dheleai, Z. Tasir, W. M. Al-Rahmi, M. A. Al-Sharafi, and A. Mydin: Modeling of Students Online Social Presence on Social Networking Sites and Academic Performance. Int. J. Emerg. Technol. https://doi.org/10.3991/ijet.v15i12.12599

[4] C. Greenhow, B. Gleason, and K. B. Staudt Willet: Social scholarship revisited: Changing scholarly practices in the age of social media. Br J Educ Technol, vol. 50, no. 3, pp. 987–1004 (2019). https://doi.org/10.1111/bjet.12772

[5] De Medio, C., Limongelli, C., Sciarrone, F., Temperini, M.: MoodleREC: A recommendation system for creating courses using the moodle e-learning platform. Computers in Human Behavior 104, 106168 (2020). https://doi.org/10.1016/j.chb.2019.106168

[6] Riyahi, M., Sohrabi, M.K.: Providing effective recommendations in discussion groups using a new hybrid recommender system based on implicit ratings and semantic similarity.

Electronic Commerce Research and Applications 40, 100938 (2020). https://doi.org/10.1016/j.elerap.2020.100938

[7] Souabi S., Retbi A., Idrissi M.K., Bennani S. (2020) Toward a Recommendation-Oriented Approach Based on Community Detection Within Social Learning Network. In: Ezziyyani M. (eds) Advanced Intelligent Systems for Sustainable Development (AI2SD'2019). AI2SD 2019. Advances in Intelligent Systems and Computing, vol 1102. Springer, Cham. https://doi.org/10.1007/978-3-030-36653-7_22

[8] AHMED HAMDI ABU ABSA, SANA'A WAFA Al-SAYEGH: E-learning Timetable Gen-erator Using Genetic Algorithms.

[9] Nebojša Gavrilović, Tatjana Šibalija: The Application Of Evolu-tionary Algorithms In E-Learning Systems. The 9th International Confer-ence on eLearning (2018).

[10] V. V. Zaporozhko and I. P. Bolodurina: A genetic-algorithm approach for forming individual educational trajectories for listeners of online courses. p. 8.

[11] Akure, Ondo State, Nige-ria, O. C. Agbonifo, et O. A. Obolo, « Genetic Algorithm-based Curriculum Sequencing Model For Personalised E-Learning System. IJMECS, vol. 10, no 5, p. 27 35 (2018). https://doi.org/10.5815/ijmecs.2018.05.04

[12] Queiroga, E.M.; Lopes, J.L.; Kappel, K.; Aguiar, M.; Araújo, R.M.; Munoz, R.; Villarroel, R.; Cechinel, C: A Learning Analytics Approach to Identify Students at Risk of Dropout: A Case Study with a Technical Distance Education Course. Appl. Sci, 10, 3998 (2020). https://doi.org/10.3390/app10113998

[13] Viloria A. and al.: Prediction Rules in E-Learning Systems Using Genetic Programming. In: Vijayakumar V., Neelanarayanan V., Rao P., Light J. (eds) Proceedings of 6th International Conference on Big Data and Cloud Computing Challenges. Smart Innovation, Systems and Technologies, vol 164. Springer, Singapore (2020). https://doi.org/10.1007/978-981-32-9889-7

[14] A. A. Alghamdi, M. A. Alanezi, and F. Khan, 'Design and Implementation of a Computer Aided Intelligent Examination System', *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 01, p. 30, Jan. 2020, https://doi.org/10.3991/ijet.v15i01.11102.

[15] Ansari, M.H., Moradi, M., NikRah, O., Kambakhsh, K.M.: CodERS: A hybrid recommender system for an E-learning system. In: 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS). IEEE, Tehran, Iran, pp. 1–5, (2016). https://doi.org/10.1109/icspis.2016.7869884

[16] Zhuhadar, L., Nasraoui, O., Wyatt, R., Romero, E.: Multi-model Ontology-Based Hybrid Recommender System in E-learning Domain. In IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. IEEE, Milan, Italy, pp. 91–95 (2009). https://doi.org/10.1109/wi-iat.2009.238

[17] Souabi Fazeli, S., Drachsler, H., Bitter-Rijpkema, M., Brouns, F., Brouns, W. van der V., Sloep, P.B.: User-Centric Evaluation of Recommender Systems in Social Learning Platforms: Accuracy is Just the Tip of the Iceberg. IEEE Trans. Learning Technol. 11, 294–306 (2018). https://doi.org/10.1109/tlt.2017.2732349

[18] A. Lambora, K. Gupta, K. Chopra: Genetic Algorithm- A Literature Review. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, p. 380 384 (2019). https://doi.org/10.1109/comitcon.2019.8862255

[19] W. Wen-jing, 'Improved Adaptive Genetic Algorithm for Course Scheduling in Colleges and Universities', *Int. J. Emerg. Technol. Learn.*, vol. 13, no. 06, p. 29, May 2018, https://doi.org/10.3991/ijet.v13i06.8442.

[20] X. Chen, X.-G. Yue, R. Y. M. Li, A. Zhumadillayeva, and R. Liu, 'Design and Application of an Improved Genetic Algorithm to a Class Scheduling System', *Int. J. Emerg. Technol. Learn.*, vol. 16, no. 01, p. 44, Jan. 2021, https://doi.org/10.3991/ijet.v16i01.18225.

## 7 Authors

**Sonia Souabi**, **Asmaâ Retbi**, **Mohammed Khalidi Idrissi** and **Samir Bennani** all work at RIME TEAM-Networking, Modelling and e-Learning Team- MASI Laboratory- Engineering.3S Research centre- Mohammadia School of Engineers (EMI)- Mohammed V University in Rabat Morocco.

Article submitted 2020-12-21. Resubmitted 2021-01-25. Final acceptance 2021-01-26. Final version published as submitted by the authors.