Use of the Data Mining Clustering Technique to Identify Student Behaviors in Virtual Environments

Felipe de Jesús Núñez Cardenas¹, Ana María Felipe Redondo², Víctor Tomás Tomás Mariano¹

¹ Universidad Autónoma del Estado de Hidalgo - Escuela Superior Huejutla.
² Universidad Tecnológica de la Huasteca Hidalguense

Abstract

Today, universities around the world are incorporating virtual platforms in order to bring their academic offer closer and increase their coverage. Under this context new schemes emerge that incorporate elements focused on meeting the needs of students in these environments. Although it is true that each student has a different learning style [1], according to different studies and models, it is also true that the development of materials within virtual learning environments must be addressed in these schemes [2]. Under this new teaching context, new questions emerge about the use that the students make of the new Information and Communication technologies that are included within the different platforms that implement the Universities for such effect [3]. This research focuses on the application of data mining algorithms, using the Clustering technique on the Weka tool, to describe the pattern of behavior of university students in virtual environments, specifically on the Moodle Platform. In this way, when knowing the different patterns of behavior, improvement projects can be implemented on the materials and activities of the virtual environment that match the learning styles predicted.

Keywords: Behavior Pattern, Clustering, Virtual Environment

1. Introduction

1.1 Virtual Learning Environments

The systems for learning management or LMS (Learning Management System) are applications oriented to the promotion of education using the internet as a means to recreate a virtual learning environment, LMS offer spaces that allow interaction between different actors of a learning environment (i.e. students, teachers, also promote individual and collaborative learning, these interactions according to the coincidence of time and space, can be classified in completely virtual (e-learnig) and mixed (b -learnig) respectively, another approach is to use it as a support for face-to-face classes [4].

There are several LMS options both proprietary and Blackboard available since 1997, as well as free mode, in this case Moodle, available since 2002. Moodle is a simple software to use and has a user-friendly graphical interface. Also, it promotes a style of constructivist and social formation. The activities that Moodle allows are: Administration, Communication, Content management, Group management, Evaluation among other activities that are recreated as it happens in a classroom environment [5].

The administration consists of operations of user management: such as registrations, modifications, elimination, management of class lists, definition of roles and control and monitoring of users access to different courses. On the other hand, it also refers to the creation, modification, visibility and elimination of the courses.

Communication allows interaction between teachers and students. Synchronous communication can be given as asynchronous, giving this communication from the teacher to the students, from the students to the teacher and from the students with students.

Content management refers to the alternatives for storing and retrieving files and operations on them, viewing them, organizing them, compressing, downloading, and so on. So that users have access to the documentation.

Group Management, is oriented to create groups and their administration, of activities, contents, photos or chats between the members of the group. The evaluation offers alternatives to create, edit and execute various types of exams, feedback, qualification and publication of qualifications, as well as statistics associated with the results obtained. All these operations are recorded through databases, user access, monitoring of activities within the platform, resources consulted, analysis of this information provides an overview of how the behavior of students and Their appreciation of educational resources and will take action in this regard, with the aim of improving the learning experience within an LMS.

1.2 Data Mining Techniques

Data mining is a field of computer science concerned with the process of attempting to discover patterns in large volumes of data sets. It uses the methods of artificial intelligence, automatic learning, statistics and database systems. The general objective of the data mining process is to extract information from a set of data and transform it into an understandable structure for later use [6].

Grouping Technique: It groups data within a number of pre-established classes or not, based on criteria of distance or similarity, so that the classes are similar to each other and different from other classes. Their use has provided
significant results in regard to classifiers or pattern recognizers, such as in system modeling. This method, due to its flexible nature, can easily be combined with other types of data mining techniques, resulting in a hybrid system [7]. A problem related to cluster analysis is the selection of factors in classification tasks, because not all variables have the same importance when grouping objects. Another problem of great importance is that currently arouses great interest is the fusion of knowledge, since there are multiple sources of information on the same subject, which do not use a homogenous categorization of objects. In order to solve these problems, it is necessary to merge the information at the time of collecting, comparing or summarizing the data [8].

1.3 Clustering: SimpleKmeans

One of the algorithms most used to make clustering is k-means (kmeans). This algorithm must first specify how many clusters are to be created, this is the parameter k, k elements are selected randomly representing the center and average of each cluster. Then each instance is assigned to the center of the nearest cluster according to the Euclidean distance that separates it from it.

For each of the clusters thus constructed, the centroid of all its instances is calculated. These centroids are taken as the new centers of their respective clusters. Finally, the whole process is repeated with the new centers of the clusters. The iteration continues until the assignment of the same examples to the same clusters is repeated, since the central points of the clusters have stabilized and will remain unchanged after each iteration. To obtain the centroids, the "mean" or "mode" is calculated according to whether they are numeric or symbolic attributes.

The SimpleKmeans algorithm is a clustering technique that gathers in tables different attributes with a similarity. It divides them into an amount of Cluster and provides the percentage of this in each of them.

2. Development

2.1 Methodology

For the present work, the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology [9] was used, it is a free distribution methodology that can work with any tool to develop any project that focuses on the implementation of data mining as shown in figure 1. The standard includes a model and a guide, structured in six phases, some of these phases are bidirectional, which means that some phases will allow to revise partially or totally the previous phases [10].

The phases are:

A) Case Study
B) Understanding of the Data
C) Preparation of the Data
D) Modeling
E) Evaluation
F) Deployment

2.2 Case Study

The University of Technology Huasteca Hidalguense (UTHH) offers educational programs of the Higher Technical University (TSU) and Engineering. Its Educational Model has implemented a b-learning scheme, in which they mix virtual environments with the real environments that allow optimizing both material and human resources, as well as undertaking strategies to support the teaching-learning process in students. The present data mining project seeks to analyze and understand the behavior taken by students of the UTHH in the use they make of the virtual platform Moodle during a period of four months.

In the UTHH through its subjects as Socio-cultural Formation and Oral and Written Expression for all the bachelors and throughout its university cycle, in which the students access from a computer equipment and do activities marked in itineraries created by the teachers of each subject, the most important activities considered for the present project are shown in the following table:
The analysis was made specifically for the students who attended as Higher University Technicians from the 5th Semester in the period January-April 2016 in the field of Oral Expression and Writing II with the purpose of having a study that determines the behaviors taken in the platform. The need for this description of behavior is due to the optimization of resources or in its effect to design new schemes that allow greater interaction with the platform. The more than 700 students from 10 different bachelors who attended the 5th Semester will be evaluated to determine their behavior during this period, which will be divided into three samples: Low, Moderate and Permanent. Below, the sample of the students is presented.

### Table 1. Sample of students from 5th Semester to analyze

<table>
<thead>
<tr>
<th>Career (TSU)</th>
<th>Groups</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Builder</td>
<td>5</td>
<td>25</td>
<td>18</td>
<td>17</td>
<td>22</td>
<td>25</td>
<td>107</td>
</tr>
<tr>
<td>Mechanical</td>
<td>5</td>
<td>21</td>
<td>24</td>
<td>28</td>
<td>32</td>
<td>30</td>
<td>135</td>
</tr>
<tr>
<td>Accountancy</td>
<td>4</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>29</td>
<td></td>
<td>122</td>
</tr>
<tr>
<td>Technology of the information and communication</td>
<td>4</td>
<td>22</td>
<td>24</td>
<td>25</td>
<td>27</td>
<td></td>
<td>98</td>
</tr>
<tr>
<td>Gastronomy</td>
<td>3</td>
<td>23</td>
<td>25</td>
<td>24</td>
<td></td>
<td></td>
<td>72</td>
</tr>
<tr>
<td>Business development</td>
<td>3</td>
<td>24</td>
<td>23</td>
<td>23</td>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td>Mechatronics</td>
<td>2</td>
<td>20</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>Process alimentary</td>
<td>1</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32</td>
</tr>
<tr>
<td>Project management and evaluation</td>
<td>1</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>44</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>719</strong></td>
</tr>
</tbody>
</table>

### 2.3 Understanding of the Data

A complete analysis of the Moodle2015 database was carried out, which contained information on all the courses and subjects that were taken in the period January-April 2016. At this stage, it was arranged to find the information that is useful within the database to be able to make a series of eventualities that give a debug panorama of the tables required for the later stages.

Among the 372 tables that the database contained, 4 had relevant information to the activities that the students performed, as well as their userid that allowed to gather them in bachelor's degree programs and groups.

In order to determine the tables to be used, an analysis of all the attributes within each table was made to determine if its content was relevant to determine the behavior of the students.

The tables used were mdl_logstore_standard_log, mdl_groups, mdl_course, and mdl_groups_members, which only important columns were used by applying join's to join them and make a comparison of them to obtain concrete data in a single table. Using the mdl_course table, the data could be grouped by bachelor's degree program and subject by means of the id and category columns.

### 2.4 Preparation of the Data

In order to obtain the information, test-and-error queries were performed, which helped to know the database and its contents by means of unions between related tables and specifications of the courses.

It was possible to obtain the tables of users per bachelor's degree program and the actions carried out during the semester.
2.4.1 Grouping by carrers.

To know the general behavior for each bachelor's degree program and to be able to make a general comparison of all of them, a minable view that allows to analyze the data, using (view) with the SQL syntax in the platform of SQLyog to maintain the data stored in a Single table and from it to be able to extract the data of userid for the comparison of tables by bachelor's degree programs.

Then, a table with different columns was created to assign the actions that each user can do.

![Tables to Analyze extracted from Moodle](image)

**Figure 3.** Tables to Analyze extracted from Moodle

```
SELECT md_logstore_standard_log.id, md_logstore_standard_log.timecreated, md_groups_members.userid, md_logstore_standard_log.action 
FROM md_groups_members 
INNER JOIN md_logstore_standard_log 
ON md_logstore_standard_log.userid = md_groups_members.userid 
INNER JOIN md_course 
ON ON_md_groups_members.groupid = md_course.id
```

**Figure 4.** Consultation to conform mined views by bachelor's degree programs

![Creation of an Action Table to Evaluate](image)

**Figure 5.** Creation of an Action Table to Evaluate
According to this database the data could be visualized as shown below:

![Figure 6. Display of the Minable View by Actions](image)

2.4.2 Separation of bachelor's degree programs by Groups

Having the information already expressed in the previous step, what was required now, was to divide into groups each bachelor's degree program, what was done was to create a view in which we wanted to insert the name (name) and the id of the user (userid) Setting as a parameter the group belonging to. Once the table that would contain the data in columns was ready, we proceed to count and insert all the data into it, for this, the same data transfer queries were generated but with different parameters.

Having the general table of students, by bachelor's degree programs and by groups of bachelor's degree programs, a file was generated with the extension (.arff), which is the format used by the Weka tool used to model the behavior of students.

![Figure 7. Creation of Minable View by Groups in a Bachelor's Degree Program](image)

2.5 Modeling

For this investigation, it was decided to use data mining tasks of the descriptive type. In particular, clustering [11] was applied to identify homogeneous subgroups within the student sample. We used the software WEKA [12]. In order to perform the analysis of the data, it was necessary to change the algorithm and the cluster to be generated, the

![Figure 8. General Arff File for Modeling in Weka](image)
choice of three clusters for this project was foreseen in the possibility of generating three groups: scarce, moderate and permanent. After that, the number of clusters and the seeds that would be used for the analysis were chosen, which would change three times for 1000, 3000 and 7000 to obtain different results and analyze them. Already configured the seeds and clusters that would carry each analysis, we proceeded to analyze the loaded arff file and its information.

Once the results were obtained from Weka, we proceeded to create a table showing the experiments performed using SimpleKmeans, which allows us to evaluate the Clusters and to know the percentage for each one of them, as well as the sum of the errors found in the process.

### Tables 2. Result of the Three Experiments with different seeds

<table>
<thead>
<tr>
<th>Kmeans</th>
<th>Cluster</th>
<th>Classified Instances</th>
<th>Sum of errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>First experiment</td>
<td>0</td>
<td>83(12%)</td>
<td>112.001</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>187(26%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>447(62%)</td>
<td></td>
</tr>
<tr>
<td>Second experiment</td>
<td>0</td>
<td>85(12%)</td>
<td>111.987</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>192(73%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>440(61%)</td>
<td></td>
</tr>
<tr>
<td>Third experiment</td>
<td>0</td>
<td>201(28%)</td>
<td>111.994</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>428(60%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>88(12%)</td>
<td></td>
</tr>
</tbody>
</table>

#### 2.6 Evaluation

In the construction of the Kmeans model, of the 717 records that formed our minable view, Weka took 100% to construct the model and 574 instances (80%) to prove it, with a precision of 92%. The algorithm was executed with 3 clusters in general and by bachelor's degree program, with three experiments each one using 1000, 3000 and 7000 seeds, the Euclidean distance with 500 iterations. In the same way, 12 attributes were used corresponding to the activities that can be performed.
The evaluation was developed on the data of training in the 10 bachelor's degree programs, represented in the following table.

### 2.7 Deployment

Within the deployment of this project we can first find the grouping by the total number of students that were analyzed to determine the behavior, this is shown in the following graph.

In the three clusters a total of 12 activities were reviewed.

Of the 717 students whose behavior was analyzed, 12%, i.e. 84 students integrate the first cluster where activities Entry Blog, Platform Access, Platform Out, Games, Submission Message, Home Course, Uploaded Task, Updated Task, Cargo File task and Visualized course, has a permanent performance, this means that they perform all scheduled activities on the platform, those weighted and reinforcing activities.

Failure Access activity has a poor performance, i.e. most students know their Access password and access correctly on the first attempt. The Games activity has a Moderate performance, i.e. half or less do not perform, this is understood because these activities are reinforcing and not weighted.

The second cluster is integrated with the performance of 182 students, that is, 26% observe a permanent performance, where the Home Course activities were carried out by the majority of the students; While the activities added Comment, Platform Access, Uploaded Task and Visualized course had a Moderate performance and activities Entry Blog, Access Failure, Exit Platform, Games, and Message sending performed poorly. Finally, in the third cluster where 62% of students, or 434 students, are concentrated, they have a permanent performance in the activities Blog Entry, Platform Access, Games, Home Course and Course Visualization. Moderated in activities Added Comment, Quit Platform, Send Message, Uploaded Task, Updated Task, Cargo File Task. Scarcity in activity Fail to access.
3. Conclusions and Future Work

Although the behavior related to the learning activities of a student in both a classroom and a virtual environment is predictable, not always those who enter a classroom obtain the same performance of the expected learning in the same way. That in a virtual environment, the performance and participation in the different tasks that make up the instructional design of a course is not homogeneous, that is, performance and learning levels vary between each of the participants, although ideally all 100%, we know that this is not possible for a number of reasons, such as learning style, study habits, affinity with the resources used, availability to self-training and self-regulation of learning, as well as availability of resources to access the platform, such as the internet, computer equipment and the burden of activities and responsibilities of each student Forward.

From the analyzed sample, it can be observed that 12% of the students have a permanent performance and 62% have a permanent and moderate performance or participation; That is to say, every 100 students, 84 of them participate actively in each one of the tasks and learning activities that have been designed within the platform, the rest has a little participation for any of the following reasons. From the results of the analysis of student behavior on the platform the following conclusions can be made.

The instructional design did not contemplate the Games activity as mandatory and only some students participated, however, the performance in this activity is Permanent and Moderate. The Blog Entry activity is Permanent, but in the Add Comment activity, the performance is Moderate, an area of opportunity is to promote not only a participation in the Blogs or Forums, but the interaction is actively promoted.

Access is logged, but the session automatically closes due to inactivity, that is why the item out of Platform is in moderate performance, this does not impact on the performance of learning activities, but rather it is a reflection of the habits of navigation. The failure of the access is given by Escaso, which is correct since most of the students access correctly in the first attempt, otherwise it is mainly because the students do not remember or they do not register their access password correctly.

Activities Uploaded Tasks, Updated Tasks, Tasks File Tasks, are affected in the results because some activities are done in team, where only one of the students sends the task and the rest of the team does not, this can be interpreted as not all Students send homework, but not necessarily so; The design of equipment within the platform is an area of opportunity, yet the performance is between Permanent and Moderate. Regarding Message activity, this is not exactly a learning activity, but rather a communication alternative, it is understood that the performance is between Permanent and Moderate because the students interact in person, minimizing their need for communication via platform; Additional media should be considered such as whatsapp groups, sms or social networks.

The activities Access Platform, Home Course and Visualization are quite obvious that they are in Permanent performance. One aspect that impacted on the sample is that the total number of students enrolled at the beginning of the school year was considered, which did not match with the total number of students attending on a regular basis, that is why the Escaso Has been increased, because of this, it is necessary to carry out a later analysis.

This analysis provides an overview of the activities students perform on the platform, offering learning resources according to their learning style is an area of opportunity.

Acknowledgements

Special acknowledgements to Ms. Elizabeth Cruz Bautista for the translation of the present work from Spanish into English

References