Data Analysis as a Tool for Optimizing Learning Management Systems

Paloma Moreno Clari  
University of Valencia  
Computing Department  
Avda. Vicente Estellés s/n  
Spain  
paloma.moreno@uv.es

Miguel Arevalillo-Herráez  
University of Valencia  
Computing Department  
Avda. Vicente Estellés s/n  
Spain  
miguel.arevalillo@uv.es

Vicente Cerverón Lleó  
University of Valencia  
Computing Department  
Avda. Vicente Estellés s/n  
Spain  
vicente.cerveron@uv.es

Abstract

The advent of the Internet has opened a scope for research in new methods and tools that may facilitate the teaching and learning processes. This has, in turn, led to the development of learning platforms to support teaching and learning activities. The market penetration of these has been such that nowadays most universities provide their academic community with some form of a learning management system (LMS).

Although a lot of effort has been put into deploying these platforms, the usage statistics that they generally provide are not generally processed to optimize their use within a specific context or institution (or confronted with quality or innovation indexes to produce useful feedback information). In this paper we propose a methodology for data analysis and apply it to the particular case of the University of Valencia (Spain). Besides, we briefly describe a computer implementation that facilitates the application of this methodology to any other existing LMS which provides usage statistics.

1. Introduction

Technological and pedagogical recent developments have caused an exponential increase on the demand for ICT (Information and Communication Technologies) based education, and a revolution in the form we understand the teaching and learning process. These changes have forced all elements involved (students, lecturers, contents and tools) to adapt to a new scenario; increased the complexity of educational programs significantly; and altered the definition of education and learning [1].

Current technologies offer almost unlimited and ubiquitous access to contents and they help improved interaction, often limited by time and space constraints in traditional classrooms. However, the provision of top ICT infrastructures and tools does not ensure the fulfillment of educational objectives [2]. It is the adequate use of these tools that results in a more efficient instruction. For this reason, plans and actions must be added to equipment and applications. These must focus on promoting and improving the quality of education [3] and on generating competences and skills among students, teachers and staff with regard to an effective ICT use [4]. Besides, they should be monitored, and their outcome formally assessed. The importance of this has been recognized within the e-learning maturity model (eMM) [5], including “Optimization” as one of its five dimensions and “Evaluation” as a process category.

Most universities offer on-campus education, and use a LMS to support teaching and learning (as opposed to pure distance learning) [6]. In this particular context, a culture for continuous improvement necessarily involves careful analysis of usage data. The results of this study can be used to improve certain aspects of the LMS, plan for new actions which have a positive educational impact, and determine the usage and acceptance level of the LMS. Furthermore, if the usage information is confronted with other variables, important relations may be discovered.

In this paper we present a methodology to perform such an analysis. The remainder of this paper has been organized as follows. First, the case of the University of Valencia and the data analysis methodology applied to this particular case are presented in section 2.
Section 3 is devoted to describe the implementation of a tool developed to facilitate the application of the same methodology to any other existing LMS which provides usage statistics. Finally, section 4 withdraws some conclusions and outlines some future research work.

2. Data analysis

The University of Valencia (www.uv.es) is one of the largest in Spain. It currently offers on-campus education to about 50,000 students in 18 schools, delivering approximately 1,500 different modules. Although by 2003 the university already counted with an important ICT infrastructure, this was not integrated under a single platform. Instead, each of a set of computer applications was used by a specific centre, department or faculty (student enrolment, data files, accountancy, etc.). In 2003, it was decided to integrate all these applications under a single LMS. After an exhaustive survey of the available platforms, the university joined the .LRN project [7] and personalized the product for the specific needs of the institution. Apart from constituting a single access point for many of the existing applications, this LMS offered a number of useful services to the academic community. In particular, the following functionalities were provided for each module: a document repository, an event calendar, a news section, the possibility of creating forums, e-mail and chat services, support for notifications, submission of homework, a repository for learning objects, an application to create Web presentations, weblogs, a photo album and a FAQs section.

After two years of usage [8], an in depth analysis of the data gathered was carried out. The objective of this analysis was first to determine the usage levels of the platform, across the two year period. Second, to obtain useful feedback that could be used to help in determining the most appropriate action plan (according to the financial and educational objectives of the university).

The entire analysis focused on three groups of measurements which were obtained for each of 18 university schools that constitute the university:

a) A number of variables related to the general usage of the platform. These include: number of lecturers, students and subjects; and a variable for the usage of each of the tools provided by the platform (e.g. document repository, forum, news, e-mail, homework submission, etc.).

b) A set of variables related to the educative innovation programs developed at the different schools within the university (e.g. number of teaching innovation programs and number of courses, coordinators, teachers and students involved in these programs).

c) Overall education quality indexes (e.g. student acceptance rate, student drop rate, number of admissions as first or second preference, student satisfaction index –measured using questionnaires–, student success index and number of months that students spend abroad under international exchange programs).

![Figure 1. An schematic view of the analysis process.](image)

Figure 1. An schematic view of the analysis process.

Note that, although the first set of variables could be obtained directly from the usage data produced by the learning platform, the second and third sets depend on other applications (some of which had already been integrated into the LMS).

All these measurements were used to derive useful information in two main ways. First a descriptive analysis used simple statistical summary measures, mainly on the first group of variables above (such as the sum, average or variance of a single variable) [9]. Then, inferential results were obtained by using multivariate data analysis techniques [10], this time combining the information on the three groups of variables. The inferential data analysis methodology is summarized in Figure 1. First, Pearson’s correlation is used as a tool to study existing relationships between the three groups of variables. After, compact variables are created, using Principal Components Analysis (PCA) on a subset of the original variables (those which have demonstrated to be more predictive or discriminative). Finally, feedback information is derived from a study of Pearson’s correlation on these newly generated compact variables. Inputs and outputs at each process stage are illustrated in Figure 2.
2.1. Descriptive analysis

In the case of the University of Valencia, statistics on the use of some of the functionalities provided evidence that the most used tool was the document repository (see Figure 3).

The document repository is a convenient tool which allows lecturers to make materials available to students on a private Web site for the module. In fact, this functionality was well accepted and quickly replaced most individual module Web sites. The information on the most used and less used tools is used to determine best candidates for improvement [11]. On the one hand, it is easy to gather information on tools which have many adepts, and improving these is a good strategy to gain new users. Besides, it is also worth analyzing the reasons behind the low usage of others.

In many cases this is due to deficiencies in training, complex interfaces or even that the user is unaware of the existence of the tool.

2.2. Initial correlation analysis

A first operation aims at reducing the number of variables involved in the analysis. From the three groups of original variables, the map of relations is studied using Pearson's correlations. The study of these values and the statistical significance associated with them makes it possible to select the set of variables that offers the most relevant information [12].

At the same time and also as a part of this first stage of the inferential analysis, dispersion graphs are used to study every pair of correlated variables. From all the variables considered that present a significant correlation, the only two which do not show a linear correlation with the rest are the number of teaching innovation programs and the number of admissions as first or second preference.

Another objective of the inferential analysis is to determine whether the usage increase between academic years is relevant in statistical terms. This has been done with the SPSS software, using the Linear General Model Module of repeated measurements ANOVA. Moreover, Huyhn and Feldt corrections and Levene’s tests have been used to verify the necessary assumptions which support the use of the statistical procedures. The former assesses the quality of variance in different population samples. The latter tests how much the sphericity assumption is violated. Besides, measurements of the effect size have been gathered. This entire process yields a quantification of the practical importance of the different variables under analysis.

The results obtained in our particular case show that the following variables have a statistical significance: number of students, number of lecturers, use of the document repository and news tools, number of teaching innovation programs, number of courses and students involved in these programs, and the student satisfaction index.

2.3. Creation of compact variables

Once a first analysis on the original variables has been performed, the next task is creating a set of compact variables for summary purposes. These new variables will gather the behavior of the original variables which have demonstrated to be more predictive and/or discriminative (according to $R^2$ values for simple and multiple regression models with $p<0.05$).
The objective is to produce a set of variables which can be easily interpreted and provide summarized information of the three themes of interest: platform usage, educative innovation and overall education quality indexes. PCA has been used for this purpose. In our particular case, this has lead to the definition of two components (compact variables) for variables related to the usage of the platform (one related to the number of users and courses and another to the use of the platform functionality) and one more for variables related to educative innovation. Regarding education quality indexes, no compact variable has been found that gathers the behavior of all statistically significant variables in this group.

2.4. Correlation analysis on compact variables

As a final step, the relationships between the summary variables obtained in the previous step are studied. This analysis will produce information on how the use of the learning platform, the educative innovation programs and the education quality indexes relate to each other. Note that this information can easily be processed to build a DSS (Decision Support System) which can be used by University managers to produce an action plan (or to predict the effect of a specific action). As an example, it will be possible to reveal important information on the impact that encouraging the use of the platform will have on student satisfaction; or the effect that increasing the number of educative innovation programs will have on the platform usage (to determine the financial investment that will be required).

Again, Pearson’s linear correlations are used. In the particular case of the University of Valencia, important relationships have been discovered. In general, the usage of the LMS is related to both educative innovation programs and the student success index.

It may also be remarked that other more surprising associations have been found. In particular, that the number of months that students spend abroad under international exchange programs predicts educative innovation during the following academic year in almost 40%. As well, the drop rate predicts the use of the platform functionality during the following year in 25%. Indeed, these points must be studied further to determine if these conclusions are genuine or just due to the short time span of the data gathered (two academic years).

Finally, linear predictive regressions of the platform usage during the last academic years have been undertaken. Although it has been concluded that in most cases the value of an indicator has a strong impact on the value of the same indicator during the following year, there is a clear exception on the use of the different tools provided by the platform. This is in part due to the actions adopted by the university in response to the first descriptive analysis. In some cases, some unused tools have been promoted and specific contents have been incorporated into existing training programs on the use of the platform.

3. Standardization of the methodology

The statistical analysis above has been described for the specific case of the University of Valencia, and thus the set of input variables related to usage coincides with those produced by the specific learning platform deployed at this institution (based on .LRN). Besides, the use of a commercial product (SPSS) and the intermediate files required during the analysis (see Figure 2) makes it difficult to automate the analysis.

A first pitfall is the input format required by SPSS. Due to the limited functionality for data import provided by this software, preparation of the input data may not be a trivial task. For this reason, we have preferred to use a XML schema [13] to define the structure of the input data. Then, a XSTL [14] transformation is used to convert the data in XML format to that required by SPSS.

Ideally, any other institution interested in the same type of analysis which has been presented in this paper would only have to provide their data according to the format defined in the XML schema. Unfortunately, it has not been possible to automate the entire analysis process by now. This is because the data analysis is sufficiently complex to require user intervention (at the stage of creation of compact variables, human interpretation is needed). For this reason, we have only automated the first part of the process. To facilitate future automation of the rest of the process we have also defined the structure of intermediate files using a XML schema.

4. Conclusions and future work

In this paper, we have presented a methodology for data analysis which makes it possible to infer useful information which can be used to enable a culture for continuous improvement. The high potential of this analysis as a prediction tool makes it especially attractive, whether as part of a Decision Support System or to discover important relations between the usage of the learning platform and other variables of interest.

The applicability of the technique has been demonstrated by using it to analyze usage data gathered by the LMS currently used at the University.
of Valencia during a two year period. Apart from determining the usage levels of the functionalities provided, the results of this investigation were used to elaborate a proposal for improvement, paying special attention to the potential impact on some education quality indexes.

In the way to standardize the methodology, we have used XML schemas to define the format for input and output data. Although the first stage of the analysis process has been automated, the complete automation of the process remains as future work.

Finally, it is our intention to replace the use of the commercial statistical software SPSS by some other open source alternative (such as R [15]). Candidates are currently under evaluation.

Completion of these two actions would result in an open source off-the-shelf product which would allow a fully automated analysis of the usage data stored by LMSs. The obvious extension to this is the development of a DSS to help university management staff in important strategic decisions.

5. Acknowledgments

This work has been partially funded by FEDER, the Spanish MEC and the Generalitat Valenciana through projects GV2008-032, TIN2006-12890, DPI2006-15542-C04-04 and Consolider Ingenio 2010 CSD2007-00018.

The authors would like to thank the personnel at the Computer Services and the European Convergence Office of the Universitat de València for their collaboration and help.

6. References


