Je-LKS
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal
PEER REVIEWED PAPERS
Vol. 13, n.2, 2017
ISSN: 1826-6223 | eISSN: 1971-8829

# A MODEL FOR USERS BEHAVIOR ANALYSIS AND FORECASTING IN MOODLE

**Mario Manzo**

IT Service Center, Naples (Italy)
mario.manzo@uniparthenope.it

**Keywords:** Moodle, User forecasting, Time series, Web performance, Distance Education.

The learning process, among its different phases, involves monitoring of users behaviour in order to extract knowledge. Details about users have significant weight to understand the interests and intentions and produce forward-looking statements, as well as keep track of the learning management system (LMS). In this work, a model to investigate and predict the behavior of users, taken to explore the additional knowledge information and predict the learning outcomes, is described. In the first instance, the information are extracted through a suitable tool, and, subsequently, are submitted to an analysis phase. Time series analysis techniques are adopted to detect partial similarities between the navigation data and, subsequently, to extract a classification. Finally, performance are measured through statistical measures to evaluate the goodness of proposed approach and test its significance. The results, obtained on Moodle platform, show that the proposed model leads to accurate outcome prediction about users behavior

for citations:
**Manzo M.** (2017), *A model for Users Behavior Analysis and Forecasting in Moodle,* Journal of e-Learning and Knowledge Society, v.13, n.2, 129-139. ISSN: 1826-6223, e-ISSN:1971-8829
*DOI: 10.20368/1971-8829/1287*

and can be adopted to improve the learning paths, both in its implementation and design.

## 1 Introduction

Despite having many potentials the web analytics technologies are poorly published. Unlike the classic scenarios, a new application field of the web analytics algorithms concerns to improve the effectiveness of distance learning. Although web analytics are rapidly being implemented in various educational settings, the current implementation indicates significant potential for the generation of knowledge, learning and education.

The strategic planning of e-learning involves planning, decision making and multiple options of implementation on different levels: faculty, curriculum and individual. Distance education can be adopted as support tool for established education systems. Otherwise, it can be implemented as an independent form of teaching, as a separate teaching program or be partially introduced. The fusion of distance and standard teaching leads to extremely interesting information.

Mining data coming from education activities involves a research field relating, but not limited to, data mining, machine learning and applied statistics to information generated by e-learning environment (Hughes & Dobbins, 2015). In this field, the aim is to improve the technique for the exploration of data in order to design new algorithms to predict the learners' performance and to improve existing contents. Learning future performance prediction can be reached through the building of a model applied to a learning environment, such as Moodle, which includes information about individual learners, behaviors, activities, knowledge and student performance. Learning environments host knowledge which can be easily recovered by extraction methods. User learning and engagement data decision-making provides a constant improvement about courses and digital environment. Afterwards, a selection of digital activities and a better organization of the materials, adopted to support student learning, can be obtained.

In this paper the problem of user behavior forecasting is addressed through a model which learns on past information. The knowledge about users is extracted, for further processing, by Google Analytics and Moodle logs. The aim is to highlight the student interaction with distance education environment and to check effectiveness of teaching activities in analyzing the learning outcome. This knowledge can be adopted in designing courses, improve the structure of course, weight assessment tasks and improve users learning experience. Finally, the quality of the prediction results are evaluated by performance measures to demonstrate the effectiveness of model in distance education contexts. The paper is organized as follows: the following sections are dedicated to related work, general architecture, preprocessing, data filtering and session definition,

knowledge extraction and users behavior forecasting. Experimental results and conclusions are, respectively, reported in the two final sections.

## 2 Related works

Artificial Intelligence technologies are capable of conferring computers the ability to reproduce some faculties of the human mind. Machines are capable of think and make decisions like humans, thanks to the intelligent exploitation of a significant amount of information and data. The applications are innumerable and transversal to all sectors including distance education. One of these concerns prediction and analysis of user behavior and is very specific and hard due to the rough nature of data. Consequently, in most cases a preprocessing phase is required. Specifically, research about the analysis of users behavior in Moodle offers several case studies, solutions and insights.

In (Cristóbal *et al.*, 2008) the application of data mining techniques, such as statistics, visualization, classification, clustering and association rule mining, on Moodle data is described. The goal is to introduce, both theoretically and practically, users interested in this new research area and in particular to online instructors and e-learning administrators.

In (Rodrigues *et al.*, 2013) the attention is focused on the importance of stress during the learning process. Stress detection in an distance education environment is an important and crucial factor to success. The estimation of the students' levels of stress in a non-invasive way is performed taking measures to deal with inclusi l'apprendimento a distanzait.

In (Horvat *et al.,* 2015) the differences in student perception of the significance of Moodle quality features and differences in student satisfaction in regard to such features are addressed. The analysis of the average waiting time for a response, feedback quality, material thoroughness, material clarity, website user-friendliness, cooperation diversity and material quantity demonstrate that the components of quality features were more important to female students.

In (Fortenbacher *et al.,* 2013) an application prototype for learning analytics, called LeMo, which collects data about learners' activities from different learning platforms is described. LeMo is a system architecture which performs user path analysis by algorithms of sequential pattern mining and visualization of learners' activities.

In (Mansur *et al.*, 2013) the behavior of students, by putting ontology on domain social learning network (Moodle), is analyzed. The activities are placed as clustering parameter according to the ontology model. The ontology is created to capture the activities of the student inside Moodle. Five attributes are adopted as group cluster for classifying the students' behavior. The constructed cluster is calculated based on the e-learning hits during the learning process.

## 3 General Architecture

The proposed architecture hosts an intelligent prediction model for users performance analysis and forecasting through different steps. Blocking approach is introduced to monitor all phases and possibly to attend in case of problems. Figure 1 illustrates an overview of the system. First, Google Analytics and logs are adopted to extract users navigation data from Moodle. The obtained data are submitted to a preprocessing stage, as described in the above section, with purpose to prepare the information for the next step. Second, users time series are diversified based on prediction to perform (related format is compatible with machine learning software adopted). Finally, prediction phase provides the forecasting users outcome. The system includes flexibility and adaptability features related to data and focuses on user interests and needs. Particulary:

- learning of user behavior;
- creation or refinement of customized study paths;
- possible attendance reporting with relative increase in hardware and software resources;

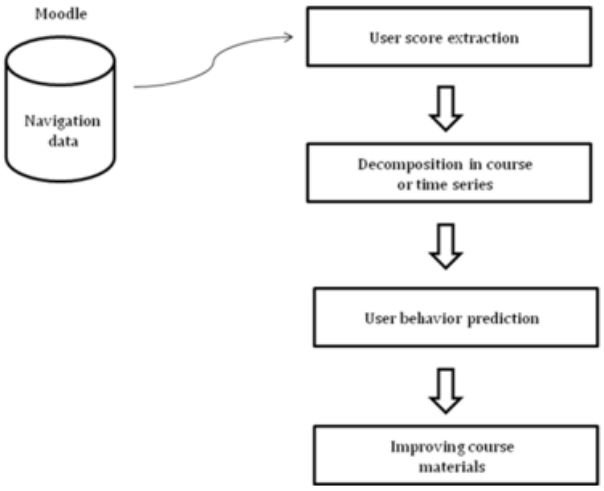


Fig. 1 - Diagram of proposed approach.

## 4 Preprocessing

In this scenario, the main problem concerns the knowledge to extract from the past in order to design educational paths and prevent less educative content. Refining content is the main goal of educators to best meet the interests and intentions of users. The first step towards this process concerns the analysis

of users navigation data in Moodle. The LMS hosts user profiles, contents, examination scores of different kinds. For this purpose, preliminary steps must be performed to prepare the contents and to obtain the data required. Rough data are extracted using a double strategy. The first concerns the extraction of Moodle logs for individual courses and platform. The second concerns data coming from Google Analytics tool as described in the next section. The choice is related to higher attendance courses from which a greater amount of data has arisen. Clearly, an integration of two data sources is perfomed to make information homogeneous and of greater quality with purpose to reach the final goal.

## 5 Data filtering and session definition

Data filtering refers to a wide range of strategies for refining data without including repetitive, irrelevant or even sensitive information. In first instance, groups representing dominant information must be isolated. Otherwise, the detection of similarities between students, activities or courses in a specific sets could be an alternative solution. Our strategy can be seen as a more top-down process, starting from a large amount of data. Sequentially, the information are refined by identifying activities with greater affluence by eliminating little indicative details. The raw data is also filtered based on the concept of session. Session means a group of interactions within a given range of time. A single session may contain multiple screens or page views, events, social interactions and transactions. A session can be considered as a container for actions taken by a user on the site. A single user can open multiple sessions, which may occur on the same day or within several days, weeks, or months. A session is defined based on two rules: timeout (after 30 minutes of inactivity), campaign change (if a user visits the site through a campaign, comes out of the site, and then returns to another campaign). In this work the timeout is adopted.

## 6 Knowledge extraction

Learning analytics includes a set of tools useful to improve learning and education. In this field, two are the main factors. First, the introduction of web counter popular today as Google Analytics. Second, the development and application of artificial intelligence techniques in distance education. Consequently, the concepts of sever logs and behavior analysis are essential. The first concerns the extraction and analysis of services monitoring, in terms of internal pages, resources, etc. While, the second provides for a set of observations useful to extract different information to follow the progress. Therefore, prediction and study of the behavior result to be more complex.

Both are used in this work. Google Analytics for navigation data extraction is adopted. It is a tool to collect and extract data about the access to websites and applications. It detects users browsing data, considering variables such as access, time, operating system, browser or location and other metrics/dimensions. Also, this information are integrated with platform log data. Given the large amount of available data produced by online activities, our attention is focused on specific subsets of information provided by Moodle. Based on this knowledge the progress is monitored and analyzed to provide improvements.

## 7 Users behavior forecasting

Forecasting involves techniques to make predictions for the future, through by information from past or present data. In the case of web analysis, prediction models are trained based on users chronology, through a single regression model, arising from a set of web pages. Time series analysis is adopted for prediction and analysis of the performance and users behavior. It concerns the application of statistical techniques to model and explain a series of time dependent data points. Time series data are composed of a natural time order, differently from machine learning applications in which each data point is a self-concept and, therefore, it is necessary to learn the layout within a data set. The main goal is to compare the prediction and real behavior of the users to improve the quality of contents. MATLAB has been chosen to analyze and manage the users navigation data. It is a high-level language for technical computing and is an interactive environment for algorithm development, data visualization and analysis. The data are organized as rows where attributes correspond to the columns. Through these available algorithms a set of prediction of the new behavior data can be performed.

## 8 Experimental Results

Experimental phase is performed on department platforms of Economic and Law of an academic institution, working through virtual architecture. LMSs host course structured in different hours of lessons, with multimedia material compliant to the SCORM (Maratea *et al.*, 2012b; 2012a; 2013). Furthermore, the platforms host more than 5000 users and, therefore, the workload, everyday, is high due to several accesses.

The building of the model, useful to learn users behavior, is a crucial step and is the starting point of forecasting. AutoRegressive Integrated Moving Average (ARIMA) (Box *et al.*, 1994) model is adopted in order to predict the behavior of users during the activities. ARIMA model acts to investigate time series having particular features and is part of the family of the non-

stationary linear processes. It starts from the assumption that the alteration in a series derived from the so-called noise. It predicts a value in a response time series as a linear combination of its own past values, past errors (named shocks or innovations), and current. Finally, it provides great flexibility and a comprehensive set of tools for univariate time series model identification, parameter estimation and forecasting. MATLAB **arima** routine is adopted. It creates model objects for stationary or unit root nonstationary linear time series model. For the experiments, the configuration includes a multiplicative seasonal model with no constant term and Gaussian innovation distribution with constant variance. In addition, model parameters are estimated through the routine **estimate**. It uses maximum likelihood based on the observed univariate time series data.

Our attention, in the following tests, is focused on number of sessions spent by users on the Moodle home page. Data are filtered in order to restrict attention to this information as Google Analytics produces data of different kind. Data are converted from CSV (Comma Separated Value) format, provided by Google Analytics, to MAT format accepted by MATLAB. Figure 2 shows results on 6 month of observation with a forecast of 20 days. The horizontal axis represents the days considered for the observation and forecast, while the vertical axis represents the sessions. As can be note, forecasting results are very close to original data. Certainly, this is a big advantage because a massive attendance can be predicted by strengthening hardware and software related to machine hosting the LMS.

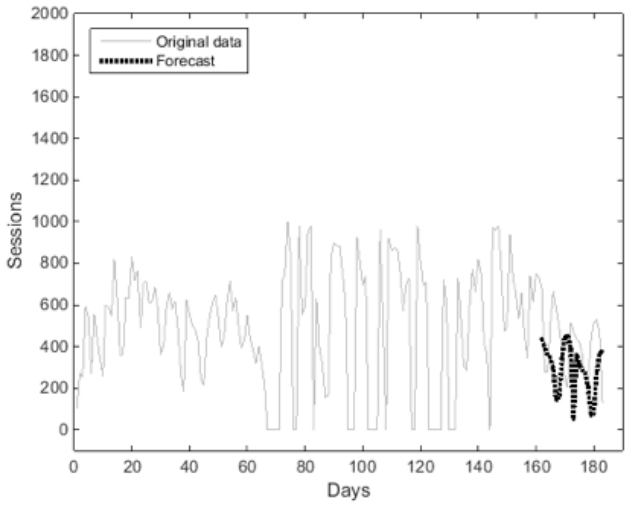


Fig. 2 - Forecasting results.

In the second test the performance are measured through by Mean Absolute Percentage Error (MAPE) between forecast $F_t$ and original data $A_t$.

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\frac{|A_t - F_t|}{|A_t|}$$

MAPE is calculated on an descending number of days of observation and a increasing number of days of forecasting (5/50, 10/45, etc). Figure 3 shows results on a 6 month of activities. The horizontal axis represents the days considered for forecasting, while the vertical axis represents MAPE values. The trend can be divided into two parts (the break point is about 50 days of forecasting). First part in which the MAPE values are low. This behavior is the result of the creation of ARIMA model on consistent data number of observation. Then the model has the best chance to learn the trend and therefore better predict. The observation is crucial for the generation of forecast values and affects, in different way, the ARIMA model. It constitutes a central point to understand and obtain information about the prior trend. Based on gained knowledge, the algorithms produce prediction values near or far respect to trend learned. Otherwise, ARIMA model for second part is not built on a large number of data. Consequently, the error appears to be higher.

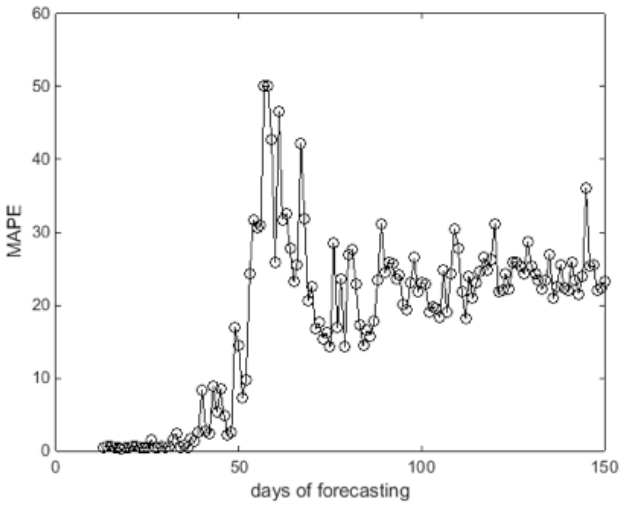


Fig. 3 - MAPE on ARIMA model.

A variant of ARIMA model, AutoRegressive Integrated Moving Average with eXogenous inputs (ARIMAX), is adopted for second testing phase. Also in this case the arima routine is adopted but with a different configuration. Figure 4 shows results on a 6 month of activities. The horizontal axis represents the days considered for forecasting, while the vertical axis represents MAPE values. Also in this case, the trend can be divided into two parts. Left to value 50 a lower error trend is achieved while an higher trend to the right. This result is due to ARIMAX model that fits the data in different way. This behavior, right to value 50, is connected to the low number of days, adopted for model building, unsuitable for accurate prediction. Clearly, the number of days is related to the educational activities held in a limited temporal slice and the building model significantly affects the performance. The alternative would be to try with a longer period that certainly cannot correspond to a real learning path.

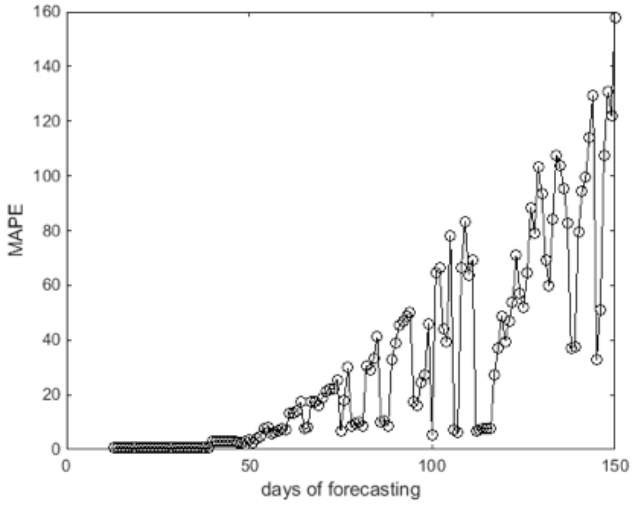


Fig. 4 - MAPE on ARIMAX model.

Last experiment is measured through by Mean Absolute Deviation (MAD). The values $e_t$ and vector E represent the error generated by the difference between forecast, producted by using ARIMA and ARIMAX models, and original data (same of previous).

$$MAD = \frac{1}{n}\sum_{t=1}^{n} |e_t - m(E)|$$

In figure 5 the horizontal axis represents the days considered for forecasting, while the vertical axis represents MAD values. The results obtained using the ARIMA model produces a greater error than the ARIMAX model especially in the range 50-70. This behavior is already detected from the previous test.

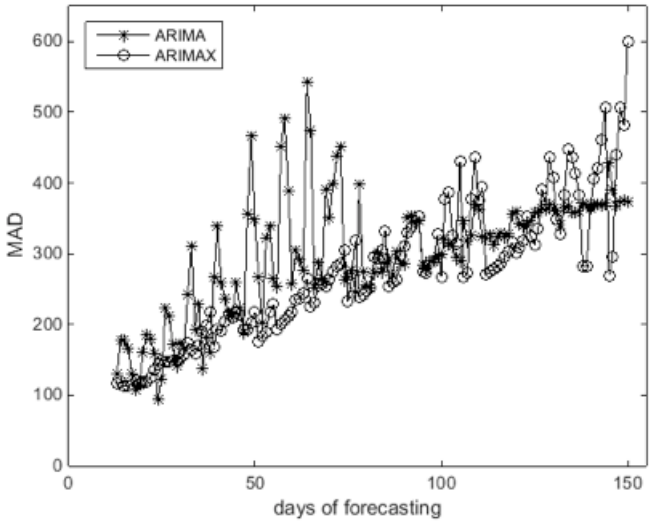


Fig. 5 - MAD on ARIMA and ARIMAX models.

## Conclusions

The organization of a blended learning strategy is not always of easy management. Starting from strong technological foundation, the goal is to determine the best content to be delivered/designed or a set of attractive activities. Accordingly, the control and monitoring of online users activities, in order to improve the design and implementation, begin essential. The extraction and analysis of LMS data became crucial in order to provide, to teachers and administrator, guidelines to monitoring users progress and action planning. In this work, an intelligent predictive model for users behavior forecasting and analysis in Moodle is presented. Forecasting analysis is a very attractive research field of recent times. Navigation data forecasting and

analysis is essential for performance accuracy of many systems to support intelligent decisions. The proposed model predicts the users behavior based on their navigation history. The goal is to verify the current trends in order to improve the content and better meet the users needs. Experimental results demonstrate the effectiveness of proposed system. Future work is in trying to analyze distributed distance education environments, individual resources, sets of learning objects and materials which require a residence time.

# REFERENCES

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994), *Time Series Analysis: Forecasting and Control*. 3rd ed. Englewood Cliffs, NJ: Prentice Hall.

Hughes, G., & Dobbins, C. (2015), *The utilization of data analysis techniques in predicting student performance in massive open online courses (MOOCs)*. Research and Practice in Technology Enhanced Learning, 10(1), pp. 1-18.

Maratea A., Petrosino A. & Manzo M. (2012a), *Automatic generation of SCORM compliant metadata for portable document format files*. In Proceedings of the 13th International Conference on Computer Systems and Technologies. ACM, pp. 360–367.

Maratea A., Petrosino A. & Manzo M. (2012b), *Integrating navigational and structural information in SCORM content aggregation modeling*. In Advanced Learning Technologies (ICALT), 2012 IEEE 12th International Conference on. IEEE, pp. 379–380.

Maratea A., Petrosino A. & Manzo M. (2013), *Generation of description metadata for video files*. In Proceedings of the 14th International Conference on Computer Systems and Technologies. ACM, pp. 262–269.

Cristóbal, R., Ventura, S. and García, E. (2008), *Data mining in course management systems: Moodle case study and tutorial*. Computers & Education, 51(1), pp. 368-384.

Rodrigues, M., *et al.* (2013), *Keystrokes and clicks: measuring stress on e-learning students*. Management Intelligent Systems, pp. 119-126.

Horvat, A., Dobrota, M., Krsmanovic, M., & Cudanov, M. (2015), *Student perception of Moodle learning management system: a satisfaction and significance analysis.* Interactive Learning Environments, 23(4), pp. 515-527.

Fortenbacher, A., *et al.* (2013), *LeMo: A learning analytics application focussing on user path analysis and interactive visualization*. In IDAACS, pp. 748-753.

Mansur, A. B. F., & Yusof, N. (2013), *Social learning network analysis model to identify learning patterns using ontology clustering techniques and meaningful learning*. Computers & Education, 63, pp. 73-86.