Web Usage Mining
Data Preprocessing and Multi Level analysis on Moodle

Nawal Sael[1], Abdelaziz Marzak[2] and Hicham Behja[3]

**[1] Laboratory of Information Technology and Modelization, Faculty of Science Ben M'sik
Casablanca, 20800, Morocco**

**[2] Laboratory of Information Technology and Modelization, Faculty of Science Ben M'sik
Casablanca, 20800, Morocco**

**[3] Laboratory of Command and Control and Production Systems, National High School of Art and Craft
Meknes, 50000, Morocco**

Abstract

This research illustrates the potential of Web Usage Mining on e-Learning domain. We use educational data mining techniques to analyze learners' behavior, to help in learning evaluation and to enhance the structure of a given course. We focus on the preprocessing task which is considered as the most crucial phase in the whole process.

Our objective is to develop a data preprocessing method applied to Moodle logs based on SCORM content structure. In earlier works [1] [2], we proposed a preprocessing tool to implement these new methods and present the first discovered knowledge.

In this research, we define new static variables according to the SCORM content tree and we apply more statistics and visualization techniques. In addition, we present multidimensional graphics in order to understand users' accesses. These aggregated variables provide to teachers and tutors interesting knowledge about students' learning process according to different levels of content accessed

Keywords: Educational data mining, Web Usage Mining, Preprocessing, Moodle, SCORM, learning process.

1. Introduction

During the last decade, numerous researches have been conducted on Moroccan higher education in order to integrate Computing Learning Environment and particularly e-Learning platforms. In this context, we have integrated an e-Learning platform in ENSAM school of Meknes[1] as learning tool, allowing students to take online courses and to benefit from the advantages of information and communication technologies.

In traditional classes, teachers are able to get information about the learning process, learners' feedback and the effectiveness of the offered learning content, and thereby the evaluation of teaching programs [3] [4].

Unlike traditional education, e-learning environments lack of teacher-learner direct relationship [5] and thus lack of information about the interaction of learners with courses and their navigations, as well as appropriate tools to monitor their activities [6]

Hence, the monitoring and evaluating learning process become difficult tasks, so teachers and tutors are obliged to find other solutions for decision-making. Helping them to understand what happens in learning sessions becomes a necessity, if not an obligation, for an efficient learning process.

The use of information technologies for the implementation of e-learning environments allows us to have a very rich source of information on the progress of learning and the how learners interact with these technologies.

Indeed, e-learning platforms are able to save all users' interactions with the environment, and can thus have information on the progress of learning as well as user profiles [7] [8]. However, information available is abundant but unstructured which imposes some essential processing like collecting, filtering, cleaning and analysis.

In this article, our aims is to analyze data generated by the e-learning platforms to assist teachers and tutors in making decisions about the course structure and its efficiency, and to provide useful knowledge about how the courses are followed by learners and help these latter in their learning process. In general, we aim to provide data mining analysis with the objective to capitalize knowledge on learners' learning process monitoring.

We focus on the first phase of the overall process of knowledge discovery from data (often referred to as: data

[1] Note : ENSAM, Site : www.ensam-umi.ac.ma

mining) applied to the Moodle platform[2]. This one is free environment with various networks of users. In addition, the majority of Moroccan universities adopt it as a tool for integration of e-Learning in the learning process.

In this paper, we present the advancements of the e-Learning research in which data mining techniques are applied in order to generate models that can be exploited to improve the feedback between teachers and learners.

This paper is organized as follows: Section 2 provides background information regarding educational data mining and its related works. Section 3 describes our proposed approach. Section 4 provides the results obtained via statistics and visualization techniques. Finally, some potential future lines of research and conclusions are outlined in Section 5.

2. Background and Related Studies

2.1 Web Mining

Web Mining focuses on the application of data mining techniques to web browsing data [9] [10]. Its techniques cover three main fields that complement each other:

- ✓ WCM: Web Content Mining or web text mining, concerned with the analysis of the web pages content,
- ✓ WSM: Web Structure Mining which deals with the structure of a website,
- ✓ WUM: Web Usage Mining, is a complex process used to extract knowledge on the characterization of Internet users attending a Web site, and the identification of their navigation patterns [11]. It consists of three phases: preprocessing, data mining and post-processing or results analysis.

2.2 Educational Data Mining (EDM)

In its official website, the international community on the EDM [3] (Educational Data Mining) defines Educational Data Mining as: an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand learners, and how they interact with those environments. This data is taken from students' use of interactive learning environments, collaborative tools, or administrative data from educational institutions and universities.

Baker and Yacef (2009) contend that the EDM can be applied mainly to improve learning style, evaluate the learning environment itself, study the efficiency of the educational support provided by the learning tools, and develop scientific research vis-à-vis learning and learners [12]. Recently, the progress of data mining techniques in education has been relatively less important than in e-commerce, although this situation is beginning to change. For there is now growing interest in the application of these techniques in the context of education[13].

Depending on their objectives, The EDM techniques are oriented towards all different actors in the e-Learning platforms, namely: learners, teachers and tutors as well as academic and administrative officials for these environments [14].

2.3 EDM Related Works

Numerous studies and applications use EDM techniques in order to give learners, teachers, tutors, and academic officials important and useful knowledge about the interaction of learners with the offered courses [13]. For example; to investigate the potential of EDM as a framework for validating the structure and design of an online learning environment [15]; to support the evaluation and validation of learning site designs [7]; to develop a tool for preprocessing e-Learning environment in order to automate selection, cleaning, structuring and enrichment tasks generally performed in the preprocessing phase [16]; to give teachers and learners tools that can help to monitor and evaluate learners' progress and the effectiveness of their learning processes [17] [18]; to personalize access to courses [19] and to provide an automatic recommendation for the learners based on the browsing history of the most active ones [20].

Romero and Ventura (2007) conducted a large review of the state of the art on EDM techniques [8]. They investigate 81 interesting researches studies published in this field between 1995 and 2005. In 2009 this review was extended and generalized when they explored 306 articles and researches studies, which they classified into 11 categories according to their objectives [14] [21].

Baker and Yacef (2009) proposed a survey on the top eight most cited papers in 2005 review and the proceedings of the EDM'08 and the EDM'09 conferences [14]. They regrouped EDM methods into five categories: prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models.

2.4 Data Preprocessing in EDM: Previous Works

Preprocessing is a tedious step in the process of web usage mining. Nonetheless, many studies propose preprocessing methods especially for the e-commerce domain [9] [22] [23]. Thus, preprocessing has begun to be used in e-learning with promising results even though the methods

[2] www.moodle.org
[3] www.educationaldatamining.org

used in this area differ from others on several levels [14] [24].

In the area of e-Learning, data preprocessing has not received sufficient analytical efforts and few studies have focused on this particular aspect. Indeed, most of the studies aiming to apply data mining techniques in e-Learning are based on the preprocessing technique employed in e-commerce. However, the context of e-Learning is very particular and differs from ordinary websites or e-commerce contexts, at the level of the structure, in the nature of the contents, or in the objectives of analysis.

Many studies focus on data preprocessing in e-Learning, for example: Koutri and al (2005) propose adapting the preprocessing methods initiated by Cooley (2003) [9] in e-commerce [25] while Marquard and al (2004) adding further constraints to these methods in order to respect e-learning context [16]: login to identify users, session is the set of user clicks to achieve a given activity, the transactions or episodes are identified through the classification of web pages on contents pages, auxiliaries pages and resources pages. However, this classification does not seem sufficient, especially with the diversity of the content offered and the frequent creation of new courses.

For Zorrilla and al (2005), a new session begins when a change in a user-course happens or when the interval between two successive transactions is more than 30 minutes [5]. However, even within the same course, the student can change the activity.

Ba-omar and al (2007) adopt 30 minutes as the maximum interval to begin new session [26].

The analysis of these different studies allows us to conclude that there are still specific tasks for data preprocessing in this domain that are not yet exploited. The majority of the data preprocessing techniques used organizes data into sessions according to course access or activity changes, but do not cover the changes that occur in the activity itself, especially for SCORM content-based activities. Further detail regarding this aspect will be provided in the present study.

3. Proposed Approach

The goals of this research are: a) to offer an analysis of e-learning environment in order to give interesting insights into the learning process, b) to assist teachers in their monitoring, and c) to help to make decision about the effectiveness of SCORM content structure and organization. The research is constructed according to web usage mining process. In [1] we proposed our new data preprocessing method applied to Moodle logs in order to structure the data and allow the analysis of SCORM content in a new way that gives more detail about learners' interaction with this content. Next, in [2] we developed our preprocessing tool and presented the first statistics gained within. In these work we apply more statistics and visualization technique, give interesting information about learners' use of course SCORM content in general and about each part of this content. We define new statistical variables and we present a multidimensional graphic to facilitate understanding learners' interactions with the environment.

3.1 Moodle Logs Analysis

Mdl_log is an unstructured table, which records any user action on Moodle. The analysis of the data available on this table and other tables describing courses and activities, allows us to conclude that action on Moodle is achieved by a particular user (login) on a given course, and in specific activity or resource.

If the user action is on SCORM content, it is made on a particular part of this content. Mdl_scorm_scoes table describes the SCORM tree of this educational content offered.

The episodes are defined according to the levels of the organization of SCORM content. Thus, we redefine a new episode concept that can be accesses to a chapter, a section or a sequence. The previous work [1] gives more detail about this preprocessing method

3.2 Preprocessing Terminology

Sael and al (2010) proposed a restructuring of the Moodle mdl_log log data based on the terminology specified below [1]:

- ✓ User: Users are defined by their login, stored in the mdl_user table and located by login id in the other tables.
- ✓ Session: All user interactions achieved between a user login and a logout.
- ✓ Visit: For a particular session, a visit is all successive accesses to the same course.
- ✓ Activity: In a single visit, it is the successive accesses to specific activity or a resource.
- ✓ Episode: If the user access to pedagogical content under SCORM, we propose a definition of the episode which can be successive user interactions in a chapter of the course, at a section in this chapter or in a particular sequence of this section.

This terminology allows us to analyze learners' interaction with Moodle courses depending on different levels of users' accesses, namely: course, chapters, sections or sequences.

3.3 Preprocessing Data Base

Preprocessed data were structured and migrated to a Relational Database, i.e. reconstruct mdl_log table. In addition to other tables describing Moodle courses and activities, we add other tables defined later (Session, Visit, Action, T_activity and Episode). These new tables are directly linked to the user action and describe his/her interaction with the learning process.

3.4 Data Collection

In this study, we used FOAD-ENSAM logs collected from Moodle database management system. The experiments were conducted with Java/C++ students in UML course proposed in this platform. This course complies with the course structure presented in [1].

Chapter→ Section → Sequence → Tasks

The course contains: 4 chapters, 14 sections, 20 sequences, 42 tasks and 3 evaluations dispatched at the end of the second and third chapters with a case study at the end of the course. The number of entries was 44507 and the number of final tracking entries was 3000.
In addition to Moodle logs, we collected demographic and final score information from Students DB[4].

4. Data Analysis and Experimentation Results

The structure of preprocessing relational DB offers flexibility and facilitates manipulating data. Its interrogation simplifies viewing, analyzing data and calculating a set of variables and indicators that describe the learning process from a different point of view

4.1 Statistics Analysis

Statistics are often the starting point for the evaluation in an e-learning system [13]. They can be generated using standard tools designed to analyze web server logs as Access Watch, analog, and Gwstat WebStat. There are also specific statistical tools to e-learning platforms [8] [27].
However, these tools do not permit to analyze certain aspects of our e-Learning Environment (SCORM). For this reason, we have developed our own tool.
In addition to simple variables presented in [2], we define new ones here in order to give more information about SCORM content uses.

Table 1 and 2 propose some calculated and aggregated variables:

Session level

Table 1: Session level variables

Variable	Description	Unit
TtSessiondt	Duration of the session	h:min
Avgsessiondt	Average session actions duration	h:min
TtSessionact	Number of total actions	number
Pacesessions	Session actions speed: number of actions divided by session duration	[actions/min]
DtFirstAccess	Date of the first access to the session	date

Chapter level

Table 2: Chapter level variables

Variable	Description	Unit
Ttchap1dt	Total Duration of the actions in chapter 1	h:min
Avgchap1dt	Average chapter 1 actions duration	h:min
Ttchp1act	Number of total actions in chapter 1	number
Pacechap1	Chapter 1 actions speed : number of actions divided by chapter 1 duration	[actions/min]
DtFirstAccess	Date of the first access to the session	date

These aggregated variables are also calculated for all other chapters (2, 3 and 4). They stem through the new data preprocessing method based on SCORM content tree. In the same way as chapters' level, we describe sections and sequences access levels.

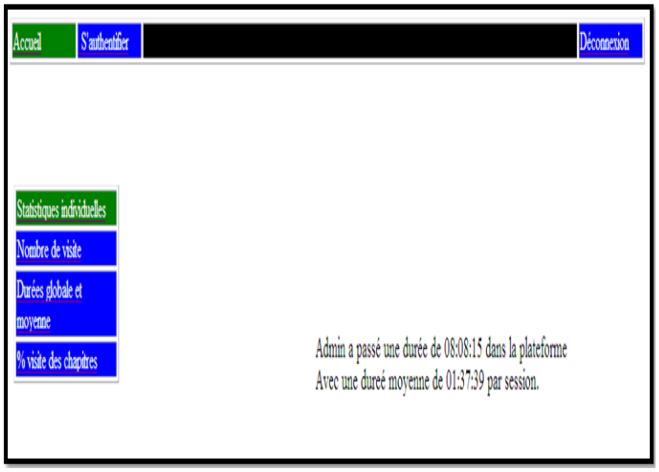
Next, in table 3 we defined other attributes that describe users' demographic information and scores:

Table 3: Users demographic attributes and score

Variable	Description	Unit
Learner_ID	The studet id	number
Gender	Learners Genders, male or female	texte
Dg_Type	Latest School type	texte
UML_score	The user score in uml module	reel
Global_score	The user global score in semester	reel
DtFirstAccess	Day of the first access in all session	date
DtEndAccess	Day of the last access in all session	date

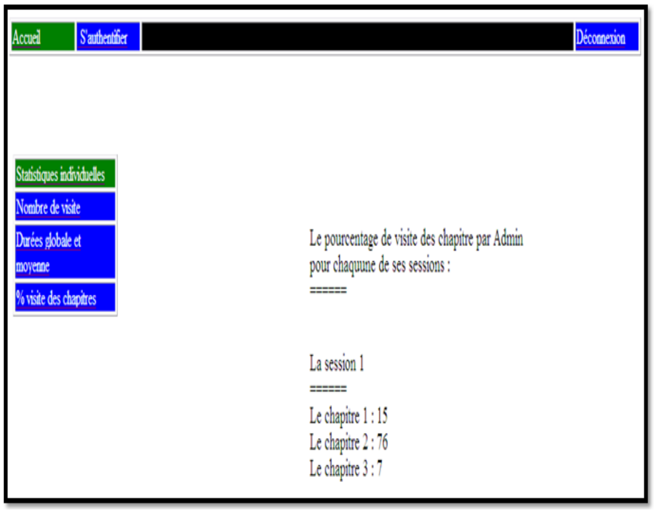
We present in "Fig. 1, 2," statistical variables calculated by our preprocessing tool which is developed in French language (first foreign language is French).

[4] School DB store all demographic information and students results



Admin a passé une durée de 08:08:15 dans la plateforme
Avec une dureé moyenne de 01:37:39 par session.

Fig. 1 Simple generated variables: Global user access duration and average sessions' duration



Le pourcentage de visite des chapitre par Admin
pour chaquune de ses sessions :
======

La session 1
======
Le chapitre 1 : 15
Le chapitre 2 : 76
Le chapitre 3 : 7

Fig. 2 Sample of aggregated data: % duration of a given learner accesses according to chapter level

4.2 Visualization

The objective of data visualization is to use graphics to enable users (learners, teachers, tutors and administrators...) to better understand and analyze large amounts of information. It allows making information less complex and more readable via the multidimensional graphics [14]. The analysis of learning progress and learners' profiles in educational content permit to understand how the content was used, how learners interact with these content over time and have suggestions on how it is designed and structured. Visualization techniques help to present calculated attributes and to analyze relationship between them in order to give more details about learners' profiles [28].

In these data analysis and visualization, we have used RapidMiner plot[5].

In "Fig. 3," we analyze the total number of done sessions for each user. We can say that there are three types of students' profiles according to the number of done sessions: less than three sessions, between 5 and 10 sessions and for many students upper than 13 sessions.

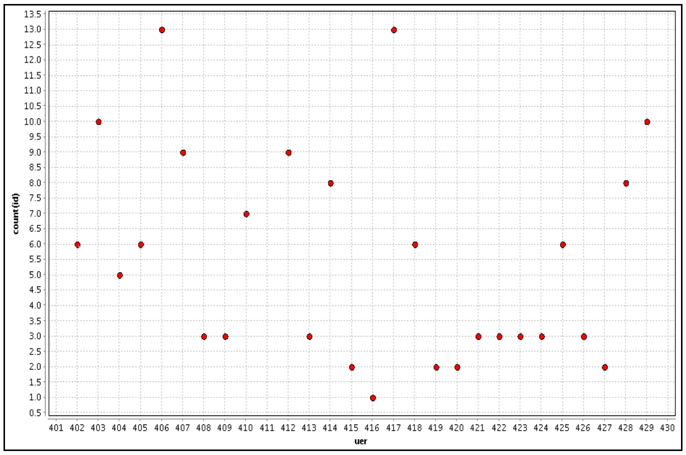


Fig. 3 Distribution of the number of session for each user.

In order to know whether the total number of sessions reflects the duration of global users' sessions, we present in "Fig. 4," the duration of all session per user.

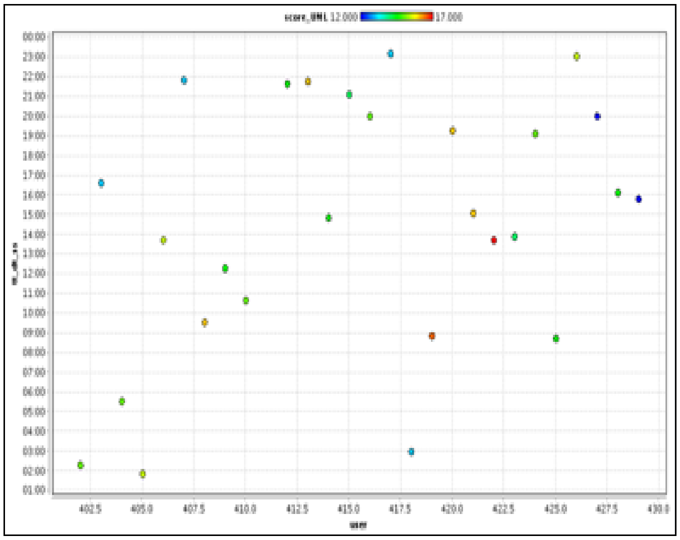


Fig. 4 Total duration of users' sessions

We notice that the distribution of learners' total session duration does not reflect the number of done sessions. We can conclude that some users open sessions and do not

[5] www.rapidminer.com

follow the content seriously (this is confirmed when calculating session pace, which reflects the browsing speed for learners).

“Fig. 4,” projects both the learners’ course score (colors), and the total sessions duration for given user. We can say that student having maximum sessions duration did not have higher score. This is more confirmed by correlation measure between score and total session duration (r^6 = -0.1772).

The distributions of users’ course usage over time “Fig. 5,” show that the total students sessions’ duration in the first day was small, increase until the seventh day, decrease between the ninth and thirteenth day and was interrupted between the nineteenth and the twenty sixth day.

This interruption can be due to the fact that student had to prepare their project in this period.

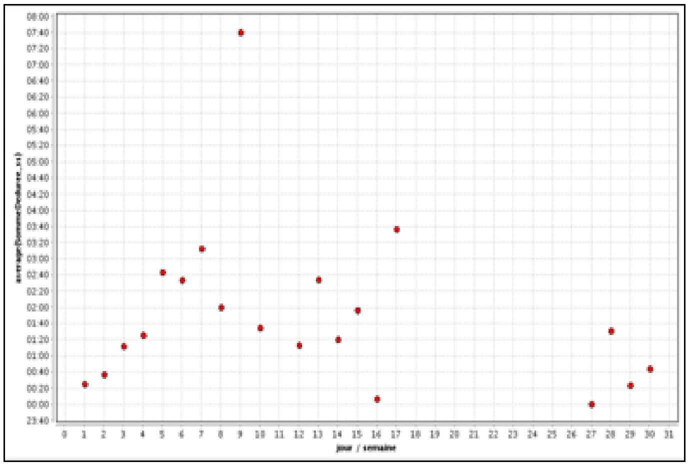
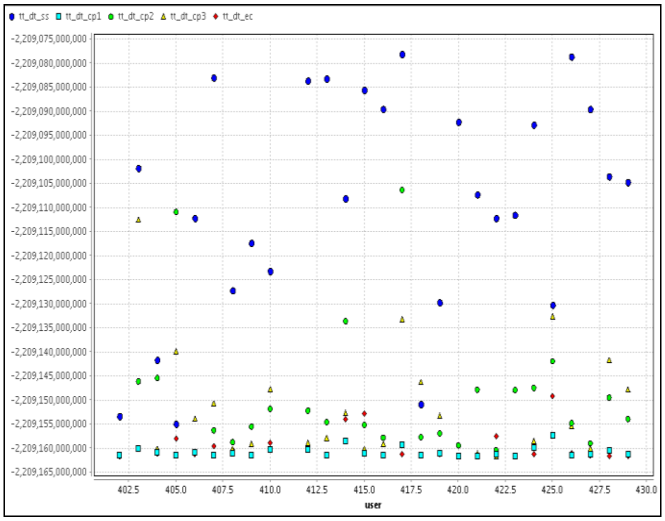


Fig. 5 Total duration of all users’ access per day

To have more details, about course usage, we analyzed the use of different part of course (chapter level). In “Fig. 6,” we calculated and presented users’ accesses duration on sessions, chapter1, chapter2, chapter3 and in the case study.



Sessions● Chapter1◘ Chapter2 ○ Chapter 3 △Case study ◆

Fig. 6 Total duration of all users accesses to SCORM content chapters (chapter level).

A phenomenon that we noticed in this results is that Chapter 2 takes more time teaching to learners (total duration is greater). However, Chapter1 and the case study were not well monitored or performed by these learners. Chapter 1 is an introduction, so it does not take much time. The second chapter is likely too long.

In “Fig. 7, 8,” we analyzed learners’ interactions with the chapters over the time. In the first days students access to all chapters perhaps in order to explore the proposed content, this interaction become more interesting in chapter 2 over time, and for the case study learners interactions were significant during the whole learning process duration.

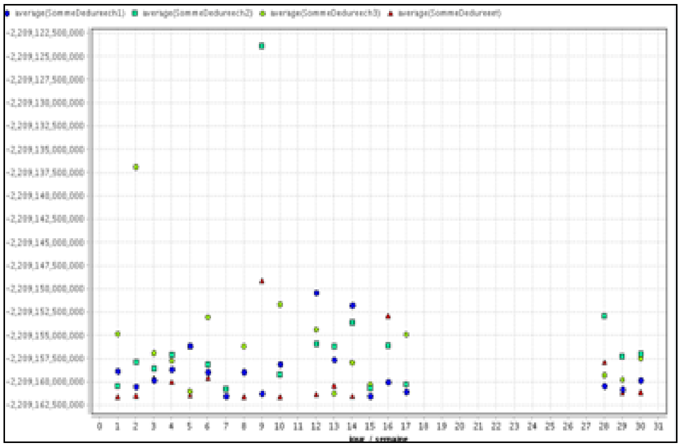


Fig. 7 Variation of students’ interaction with chapters per day

[6] r : correlation coefficient, Pearson product-moment correlation coefficient

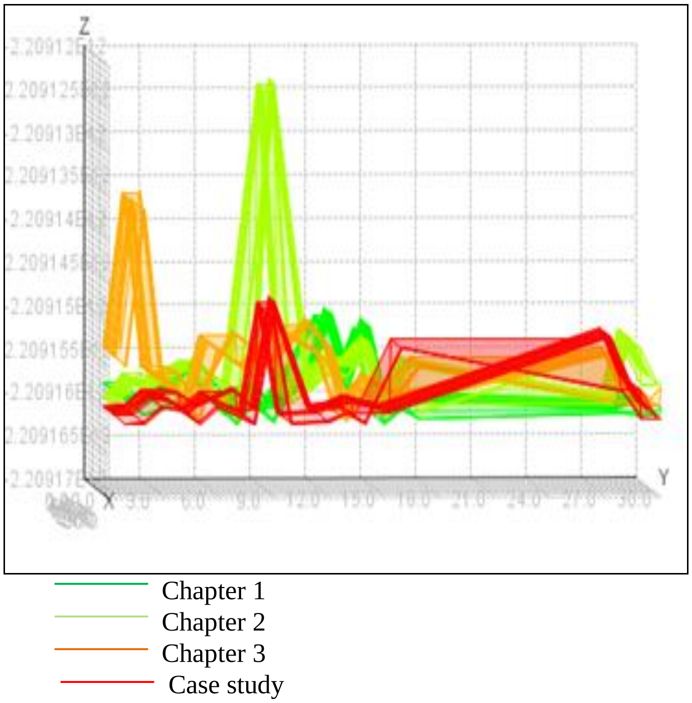


Fig. 8 Variation of students' interaction with chapters per day in 3D plot

In "Fig. 9," we analyzed learners' interactions with the case study and its relationship with the final score.

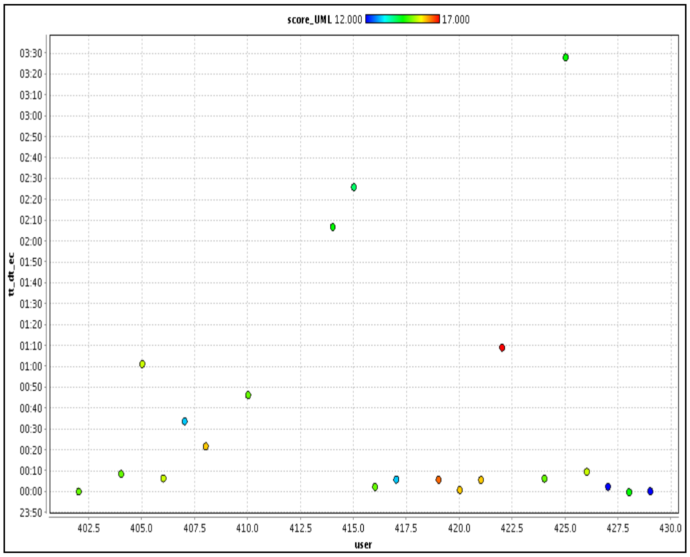


Fig. 9 Total duration of users accesses in the case study per user and final score.

We deduce that most of students' accesses duration in the case study was less than 10 min. The course score for the students who spent more time in this part of content is good. Nevertheless, it seems it does not affect directly the score of all learners.

An iterative analysis in the same way on the learners interactions with different sections for given chapter can give us more details about the parts of the SCORM content which might be more difficult or possibly not well structured. For example, we can divide chapter 2 into two parts (the most time-consuming chapter), to keep users concentration on.

Combining these observations with the scores of learners during the evaluations related to each chapter (not the final score). We can have more details about the interaction of users with this course with the aim of contribute greatly to the evaluation of its effectiveness and the validation of its current structure.

5. Conclusion and Future Works

In this research, we presented a new WUM research for e-Learning Moodle platform. The study covers four stages, two of them were exposed throughout this article alongside the results.

We deduced that even if the preprocessing phase is a crucial step in the overall process of web usage mining, its achievement is a key element in the success of this process. This research aims to help firstly the teachers and the tutors and to facilitate their tasks by automating the preprocessing steps and giving important knowledge on the behavior of learners and their interactions with the course SCORM content.

Our first objective is to allow them to know all users' interactions with course content and facilitate those interactions analysis via graphical representations that are simple and easy to analyze.

In the next steps, we plan to apply clustering techniques in order to analyze the structure of the group of learners and association rules mining to find more relationship between different parts of the offered SCORM content. We will then analyze the results obtained from these steps to help teachers to evaluate the effectiveness and structure of this educational content, and to offer an automatic recommendations for teachers/tutors and learners.

In the future, we suggest generalizing this research to analyze all the platform courses and develop a WUM tool that analyzes deeply different activities.

References

[1] N. Sael, A. Marzak, and H. Behja, "Prétraitement avancé des fichiers logs Pour une plate forme d'enseignement à distance" in NGN2010. 244-248.

[2] N. Sael, A. Marzak, and H. Behja, "Investigating an Advanced Approach to Data Preprocessing in Moodle

Platform" International Review on Computers and Software. Vol. 7 N. 3: pp. 977-982

[3] J. Sheard, S. Ramakrishnan, and J. Miller, "Modelling Learner and Educator Interactions in an Electronic Learning Community" , Australian Journal of Educational Technology, 2003, 19(2), pp. 211-226.

[4] J.-M. De Ketele and X. Roegiers, Méthodologie du recueil d'informations, fondements des méthodes d'observation de questionnaires d'interviews et d'étude de documents. De Boeck, 2009.

[5] M Zorrilla, S. Mill´an, E. Menasalvas. "Data webhouse to support web intelligence in e-learning Environments". In: Proc. of the IEEE International Conference on Granular Computing, Beijing, China, 2005, GrC pp. 722-727.

[6] R. Hijon, A. Velazquez, "E-learning platforms analysis and development of students tracking functionality". In Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, June 2006, pp. 2823-2828.

[7] M.-E. Zorrilla, D. Mar´ın, E. Alvarez, "Towards virtual course evaluation using Web Intelligence". In: Moreno D´ıaz, R., Pichler, F., Quesada Arencibia, A. (eds.) EUROCAST, Springer, Heidelberg 2007. LNCS, Vol. 4739, pp. 392–399.

[8] C. Romero, and S. Ventura, "Educational Data Mining: a Survey from 1995 to 2005". Expert Systems with Applications, 2007, Vol 33(1), p.135-146.

[9] R. Cooley, "The Use of Web Structure and Content to Identify Subjectively Interesting Web Usage Patterns". ACM Transactions on Internet Technology, 2003, Vol 3(2), pp. 93-116.

[10] L. Bing, "Web Data Mining: Exploring Hyperlinks, Contents and Usage Data", Springer, 2011.

[11] B. Mobasher, O. Nasraoui, B. Liu, B-Eds. Masand, "Web Mining and Web Usage Analysis 2004" - revised papers from 6 th workshop on Knowledge Discovery on the Web, , Springer Lecture Notes in Artificial Intelligence.

[12] R. Baker, and K. Yacef, "The state of educational data mining in 2009: A review and future visions", J. Educ. Data Mining, 2009, Vol. 1, no. 1, pp. 3–17.

[13] C. Romero, S. Ventura, M. Pechenizkiy and R. Baker, "Handbook of Educational Data Mining". New York: Taylor & Francis, 2009.

[14] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art". IEEE Transactions on Systems Man and Cybernetics Part C. Applications and Reviews, 2010, Vol 40(6), pp.601-618.

[15] L.-D. Machado, K. Becker, " Distance Education: A Web Usage Mining Case Study for the Evaluation of Learning Sites", ICALT, 2003, pp. 360-361.

[16] C.-G. Marquardt, K. Becker, D. Dubugras and A. Ruiz. "A Pre-Processing Tool for Web Usage Mining in the Distance Education Domain", IDEAS 2004, 2004 pp.78-87

[17] A. Merceron, and K. Yacef, "Educational Data Mining: a Case Study" . AIED, 2005, pp. 467-474.

[18] E. García, C. Romero, S. Ventura and D-C. Castro, "An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering". User Model. User-Adapt. Interact, 2009, Vol 19(1-2), pp.99-132.

[19] M-K. Khribi, M. Jemni, O. Nasraoui, "Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval". ICALT 2008, pp.241-245

[20] C. Romero, S. Ventura, A. Zafra, and P. De Bra, "Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems". Computers & Education, 2009, Vol 53(3), pp.828-840.

[21] Sfenrianto, H. Suhartanto, Z-A. Hasibuan, "A dynamic personalization in e-learning process based on triple-factor architecture," Computing Technology and Information Management (ICCM), 2012 8th International Conference on , vol.1, pp.69-75, 24-26.

[22] D. Tanasa, and B. Trousse, " Advanced Data Preprocessing for Intersites Web Usage Mining". IEEE Intelligent Systems, 2004, Vol 19(2), pp.59,65.

[23] N. Sael, A. Marzak, and H. Behja, « Web Usage Mining, Proposition d'une démarche pour le prétraitement des fichiers logs ». WOTIC 2009.

[24] J-L. Hung, K. Rice, A. Saba, " An educational data mining model for online teaching and learning". Journal of Educational Technology Development and Exchange, 2012, Vol 5(2), pp.77-94.

[25] M. Koutri, N. Avouris and S. Daskalaki, " Chapter 7: A survey on web usage mining techniques for web-based adaptive hypermedia systems" , in S. Y. Chen and G. D. Magoulas (ed), Adaptable and Adaptive Hypermedia Systems, IRM Press, Hershey, 2005, pp.125-149.

[26] H. Ba-Omar, I. Petrounias, F. Anwar, " A Framework for Using Web Usage Mining to Personalise E-learning". ICALT 2007, 2007, pp. 937-938.

[27] M. Goyal, R. Vohra, "Applications of Data Mining in Higher Education", in International Journal of Computer Science Issues, 2012, Vol. 9 Issue: 2, pp. 113-120.