Organised Access to Historical Student Data

Project proposal

Zeqi Fu
University of Adelaide
Adelaide, South Australia
5005 Australia
zeqi.fu@student.adelaide.edu.au

Minh Tam Phan
University of Adelaide
Adelaide, South Australia
5005 Australia
minhtam.phan@student.adelaide.edu.au

Zeyu Lin
University of Adelaide
Adelaide, South Australia
5005 Australia
zeyu.lin@student.adelaide.edu.au

1 Introduction and Motivation

The school of Computer Science regularly analyses data collected as part of teaching to improve teaching practice and to identify better ways to assess and evaluate student work. Over the years, huge amounts of data have been collected and the problem now is organizing it for searching and trend analysis.

We have two data sources: the web submission system(including old and new data systems in file directory structure) and the Moodle forum system (see Fig. 1). Consider the following question: After an assignment is released, when do students begin to download resources and when do they submit? The former relates to user activities in Moodle forum while the latter relates to activities in web submission. User activities of the Web Submission system is not tied to the Moodle forum system because they are separate systems. So currently there is no way to answer this question.

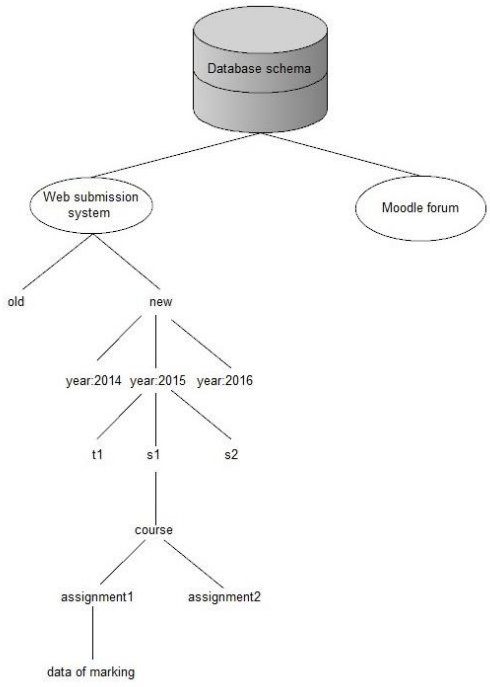

Figure 1: Two data sources of our project.

What we are doing is to tie the two systems together by designing a database schema that will take into consideration and incorporate both systems (see Fig. 1). The database should be structured in such a way that it has the capability to answer a series of questions that reporting functions of most up-to-date learning management systems cannot tackle. Here are some more examples of questions (see Table 1).

Table 1: Example questions that the database schema will answer

1. List all the students who over the years have only worked a couple of days around the assignment hand in points.
2. Activities of students in the duration of assignment.
3. Show students behaviors with GPA higher than 6.
4. Do students hand in the first assignment on time or late?
5. Show all the students who submit this assignment after this date.
6. All the events of users in a given time
7. Course enrolment time and unenrolment time distribution
8. Assignment submission time distribution.
9. Number of students' course accesses on a date.
10. Number of accesses on single resources.
11. Overview forum activities of all students.
12. Are students busy on Monday?

Why are these kinds of questions important? Because the answers to these questions will allow us to know the student behavior patterns. This will help academics identify at-risk students who are in danger of failing or not completing a course. This will help figure out the relationship between marks and student behaviors. This will also help academics know the frequency and time distribution of students' accesses to courses, resources and so on. In a nutshell, the database schema will have the capability to answer a series of questions. The answers to these questions will improve teaching practice and identify better ways to assess and evaluate student work.

In consideration of ethical issues and privacy, all data will be anonymized and will be identified in a way that allows us to associate actions without identifying the students. In order to import the current data into the schema we design, we need to write extraction and insertion scripts that will work with the file-based storage for the current data.

We will also provide a GUI-front end to allow a member of staff to query the database, extract data, aggregate it, and export it. Login function will be provided to allow teachers or researchers to use the application. Different kinds of charts (e.g. enrolment time distribution, number of accesses on single resources) and their corresponding reports will be provided.

The potential of the work is that it can be put into real use. Lecturers and researchers in the school of Computer Science can use the application to know more about students' behaviors and improve teaching practice. Ideally, the database schema should be extensible. In the future, we might add more data sources (e.g. student GPA) and expand the schema to allow it to answer more complicated questions. Also, we are trying to design a schema that will not only be suitable for Moodle, but also for other online learning management systems and other learning softwares.

2 Background

Online learning management systems (LMS) are web applications that assist lecturers with arranging and administrating learning environments. LMSs are becoming increasingly popular in modern education [1]. LMSs generally provide lecturers with records of system and students' behaviors and some reporting functions. Analyzing LMS logs in view of the course and students' activities is essential to change and perfect courses [2].

Edwards et al. investigated on student submission logs and concluded that students beginning to do the assignment early (which is recorded by the time of first submission) have greater chances of doing better in the assignment [3]. They also mentioned in their study that students who have good or poor academic performance all the time may show analogous behaviors, e.g. good students may begin the work late and can work under stress.

Nandi et al. measured engagement through student participation in forums, which was also taken as a mark predictor. The outcome indicates that students having a better mark participate more in forums [4]. However, it is noteworthy that the analyzed course was totally online and contacts between students and lecturers were completely through the forums, thus resulting in higher forum participation. The results will not apply to courses not totally online where students contact lecturers using other means instead of the forum, thus resulting in lower forum participation.

Ceddia et al. tracked student behaviors through an online learning website. Web logs were used to carry out tracking and student behaviors are classified as purposeful or browsing behavior. It turned out that with course progressing there are more purposeful behaviors and less browsing behaviors [5].

These studies above give some fruitful ideas of analyzing data of learning management systems. However, they only focus on some particular aspects of student behaviors. Data collection and data analysis in these studies are not in a structured and orgranized way.

Moodle is an open-source and free online learning management software. Moodle is applied to hybrid learning, remote teaching and other online education projects in colleges, educational institutions and other departments. It is flexible and customizable. It can be applied to set up systems with online courses for teachers and lecturers to achieve educational targets [6,7].

There are a number of researchers who applied modern data mining techniques to analyze Moodle data [2]. Knowledge and pattern behind Moodle data can be discovered and extracted.

Castro, Félix, et al. provide an overview of the use of various data mining technologies (Genetic Algorithms, Neural Networks, Visualization and Clustering Methods, Intelligent agents, Inductive Reasoning and Fuzzy Logic) in e-learning [8].

Romero, Cristóbal et al. conduct an investigation of the use of data mining technologies in online learning management softwares [9]. Data of Moodle is used for analysis. The whole process for mining Moodle data is introduced in detail. Also, information regarding how to use major data mining technologies, including classification, statistics, clustering, visualization, and association rule are provided.

Romero, Cristóbal et al. compare different kinds of data mining technologies in the application of the categorization of students according to their Moodle activities data and the grades gained in the courses [10]. A particular tool integrating the whole process of data mining is provided to lecturers for their convenience.

Most of these studies above pay attention to complex and specialized data mining techniques which turn out to be incomprehensible to the general academic audience.

There are a number of learning analytics tools related to Moodle which have already been presented.

Zhang, Hangjin et al. presents students' learning activity tracking by exploring Moodle logs [12]. The tool Moodog is provided, which allows lecturers know how students are using course related resources and allows students to compare their own performance with their peer students in the class.

Mazza, Riccardo et al. proposes a visualization tool for Moodle, which is called GISMO. Moodle logs are exploited to generate graphs, which can be used by lecturers to track and analyze student activities [14]. GISMO is presented as a module of Moodle, which can only be seen by lecturers. Lecturers can know more about the class as a whole and analyze the situation of the whole class.

Mazza, Riccardo et al. proposes the MOCLog project in order to analyze learning related activities in view of both the learning process and the results [13]. MOCLog is based on GISMO. It has reused some of GISMO's major components. Different users can be recognized and different reports will be provided to them according to their role.

Dierenfeld, Helena et al. proposes a tool called Excel Pivot Tables. It can be used to generate learning statistics from Moodle [15]. Moodle logs can be exported in excel format, which can then be used to generate Pivot Tables. By using this tool, users can handle huge amounts of data at ease and quickly, summarizing

underlying crucial information from the data and perform complicated calculations quickly.

Sampayo, F. C. proposes a tool called Analytics and Recommendations. It is incorporated in Moodle as a module, which could be accessed by lecturers and students [16]. It can visualize students' participation in different activities of a course. It can also make some recommendations on activities to students for the purpose of improving their academic performance. Colorful tables and graphs are provided by this tool.

All the above tools support visualization, which facilitates the analysis of student activities. However, they are all based on Moodle and they do not support multiple data sources other than Moodle.

There are also some other works related with analyzing Moodle logs. Konstantinidis, Andreas et al. adopts a new method (using excel macros) to present data from Moodle logs [17]. Casany Guerrero et al. analyzes Moodle logs in order to know the types of device(computer or mobile) students use to view the course. This information will help figure out which part of the original system should be remade for mobile devices [18].

There is only one study which focuses on designing a database schema to structure data. Krüger Andr é et al. propose a data model to organize data that most learning management softwares normally store in scattered places into a unified schema [11]. The data model is intended for automating the time-consuming data preprocessing step and accelerating the data exploration step that comes before data mining. An implementation for Moodle data is presented. Analysis has been carried out on a real course. This study has similarity with our work. However, it does not present a complete software solution. It does not combine several different data sources. It does not provide an application which also provides login and visualization of data by using charts.

To sum up, all the previous studies mainly fall into 4 categories. The first type of studies analyze some data of learning management systems. They only focus on some particular questions and aspects of student behaviors. Data collection and data analysis in these studies are not in a structured and orgranized way. The second type of studies mainly focus on applying modern data mining technologies to analyze Moodle logs. The complex and specialized data mining techniques in these studies are incomprehensible to the general academic audience. The third type of studies propose learning analytics tools related with Moodle. These tools support visualization, which facilitates the analysis of student activities. However, they are all based on Moodle and they do not support multiple data sources other than Moodle. The fourth type of study bears similarity with our work. The study proposes a data model to organize data that most learning management softwares normally store in scattered places into a unified schema. However, it does not present a complete software solution. It does not combine several different data sources. It does not provide an application which also provides login and visualization of data by using charts.

In view of the strengths and shortcomings of all the previous work mentioned above, our work will focus on designing a database schema that takes into consideration multiple data sources and has the capability to answer a series of complicated questions related with learning behaviors. We will provide an application which is easy to use. Complex processes such as data processing and data analysis are done in the background, being transparent to users. The application will provide login function and data visualization by using charts. Users will be allowed to query the database, extract data, aggregate it, and export it.

3 Architecture

The architecture diagram of the system is below (see Fig. 2).

The architecture of the system applies the 3-tier model including presentation, business logic and data access layers. By separating the system into different modules, 3-tier application architecture provides a flexible, reusable and upgraded model. Apart from the usual advantages of modular software with well-defined interfaces, the 3-tier architecture is designed for the purpose of decoupling the components independently in response to changes in requirements or technologies. For example, a change of the interface at the presentation layer would only affect the user interface code.

Users manipulate on the website at the presentation layer, established by PHP programming language. In fact, PHP, which is a scripting language, is widely used and particularly suited for web development and can be embedded into HTML. One difference that distinguishes PHP from other client side like JavaScript is its basis of execution on the server. Furthermore, the best things in using PHP is definitely easy to be used and it offers numeric advanced features for professional programmers.

The third-party library which is selected for generating reports is JasperReports at the presentation layer. Fundamentally, Jasper-Reports Library is recently considered as one of the most popular open source reporting engines. In addition, it is totally written in Java. Apart from being able to consume data coming from any kind of data source, it can produce documents that can be viewed, printed or exported in a variety of document formats such as HTML, PDF, Excel, OpenOffice and Word.

In business logic layer, all essential business objects and web services are created for handling the requests coming from the presentation tier. Web services are basically deployed on web server; therefore, they can be invoked remotely and easily. The requests from users through actions on the website are then transferred to the business logic layer as requests to the web services. Each web service is represented as a class containing web methods to handle the requests. At the business logics layer, web services are built based on JAX-WS technology. The Java API for XML Web Services (JAX-WS) is a Java programming language API for creating web services, particularly SOAP services. JAX-WS is one of the Java XML programming APIs. It is part of the Java EE platform. Each method of the service interacts with the method at data access layer to manipulate CRUD actions on database. Data access layer is a layer of a computer program which provides simplified access to data stored in persistent storage of some kind, such as an entity-relational database.

Hibernate ORM, used at data access layer, is concerned with data

persistence as it applies to relational databases (RDBMS). The objects at business logic layer are mapped consistently with those at data access layer through Hibernate ORM. In addition, all methods which are created at this layer will directly interact with the relational database. Hibernate is an implementation of the Java Persistence API(JPA) specification. As such, it can be easily used in any environment supporting JPA including Java SE applications, Java EE application servers, Enterprise OSGi containers, etc. Furthermore, Hibernate is considered a strong framework because of high performance, scalability, reliability and extensibility.

The database of this project is managed by MySQL, which is an open-source relational database management system(RDBMS). The major advantage of MySQL is that it is free and usually available on shared hosting packages and can be easily set up in a Linux, Unix or Windows environment. If a web application requires more than database, requires load balancing or sharding, it is easy to set up maybe instances of the database requiring only the hardware costs, as opposed to commercial databases that would require a single license for each instance. The tables of database contain data which are mainly extracted from two data sources: WebSubmission and Moodle.

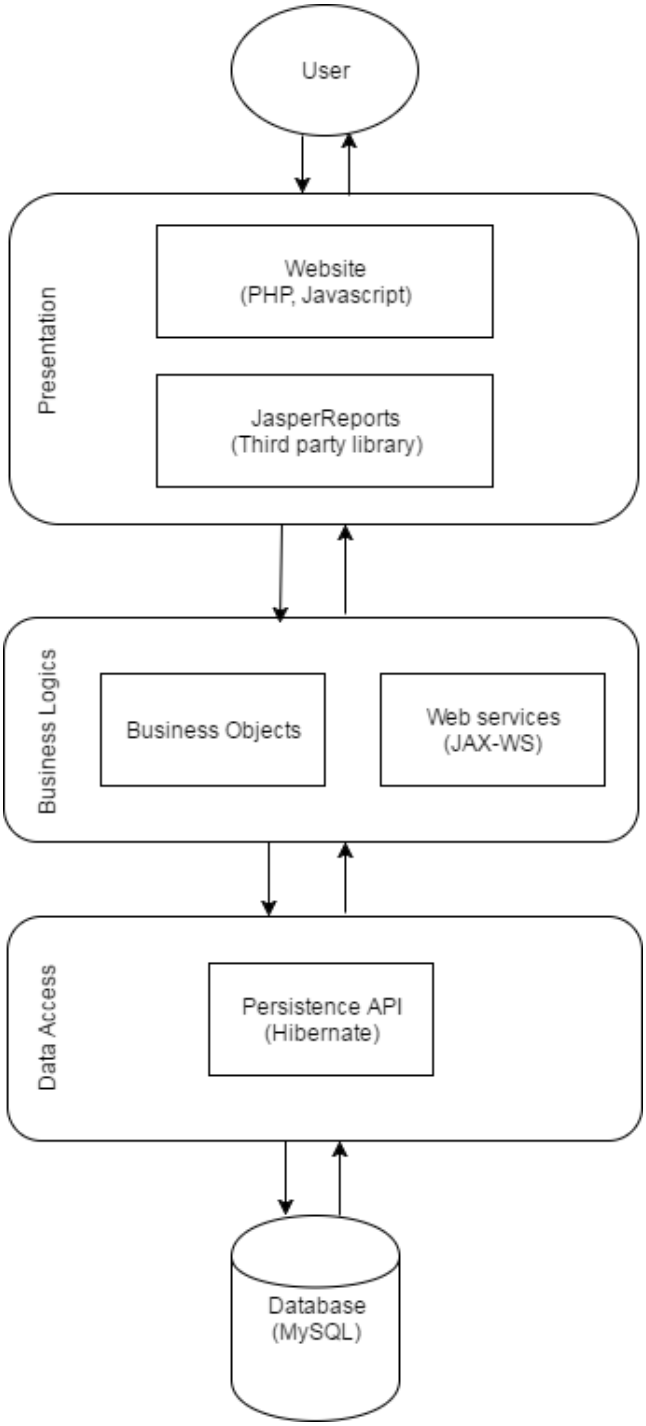

Figure 2: System architecture.

4 Research Methodology

The research method we use is survey research. We will evaluate whether what we built is useful and better than the current tools. Other research methods are not suitable for the evaluation of our work. A controlled experiment explores the effect of changes of

one or more independent variables on one or more dependent variables, enabling us to know exactly the relationship between the variables and if there is a cause-effect relationship between the variables. Controlled experiment is not suitable because we don't have variables. Case studies allow us to know the reason and the way that certain phenomena happen. Ethnography focuses on observing and studying a group of people for the purpose of understanding their social behaviors. It is obvious that controlled experiment, case studies and ethnography are not suitable. Action Research focuses on solving a real-world problem while at the same time learning from the process and experience of solving the problem [19]. The way we do the research project bears similarity with action research. We put forward a plan, get feedback from clients and reflect upon it, then we improve and modify the application. There are several such cycles. However, action research is not suitable for evaluating whether what we built is useful and better than current tools.

There are generally two types of surveys: Questionnaires and Interviews.

Many advantages of using questionnaire can be seen. It's easy and not costly to manage. Same materials can be sent to lots of people. Respondents can do the questionnaire when they are available. However, there are also some disadvantages. Response rates are usually low. It may not be suitable to ask for lengthy and detailed writing [20].

Interview is different from questionnaire. There are some advantages. It allows the interviewer to investigate further or ask follow-up questions. It's much easier when we want viewpoints or thoughts from respondents. However, the disadvantages are also noteworthy. Interviews take huge amounts of time and require a large number of resources. Interviewers have to be trained well beforehand.

We will use questionnaire because it's cost-effective. There is not much difficulty in administration. Our research focuses on evaluating whether what we built is useful and better than the current tools. We don't need many open-ended questions and opinions from the respondents because of the nature of our research.

The population for the survey is all the academic staffs and tutors from different schools and faculties in the University of Adelaide. The population size is around 3000. The ideal sample size for the survey is 341 in order to get a confidence level of 95% and a confidence interval of 5. The sampling method we will use is stratified sampling. Each school is considered as a strata. Random sampling is used within each strata. The sample size of each strata is proportionate to the population size of that strata. For example, if there are in 50 lecturers and tutors in the school of computer science, the sample size for the school of computer science will be 6 (50*341/3000). The reason for choosing stratified sampling is that it increases the sample representativeness through reducing sampling error. Also, in order to improve the response rate, we will offer some inducements (e.g. lottery).

In order to carry out the survey. We are going to complete the following steps.

(1) Design the questionnaire

There are two parts in our questionnaire. The first part includes compulsory closed-ended questions (see Fig. 3). They are questions based on level of measurement which will thus have structured response formats [21]. These questions focus on evaluating whether what we built is useful and better than the current tools. The option "I don't know" is provided to reduce the possible bias. The baseline for comparison is the Moodle reporting facilities and other learning analytics tools. We will provide links to the demo, screenshots or introduction of these tools in our survey. The second part incorporates some voluntary open-ended questions (see Fig. 4). The answers to these questions will help improve our application.

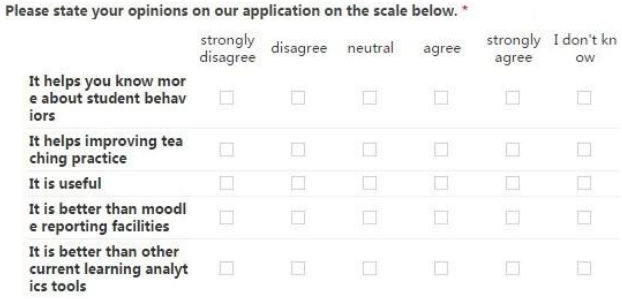

Figure 3: First part of the questionnaire.

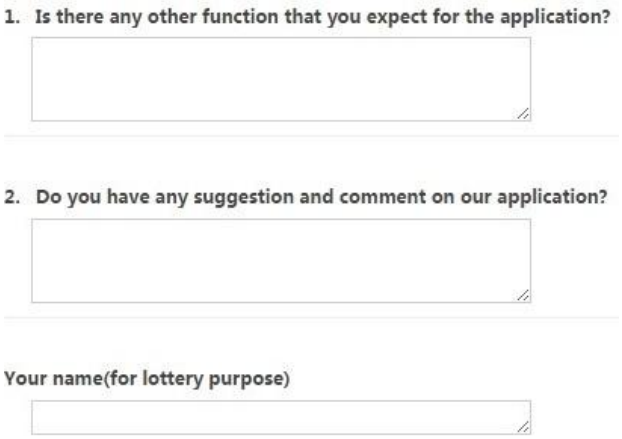

Figure 4: Second part of the questionnaire

(2) Send out the questionnaire

We will send emails to 341 academic staffs and tutors. The reason for choosing email is that it's not costly and easy to administer. The email will include the link to our application, the link to the questionnaire and the links to the demo, screenshots or introduction of Moodle reporting facilities and other current tools. Brief introduction will be given on our application and the purpose of the questionnaire. Also, the inducement for the survey will be highlighted in order to increase the response rate.

(3) Analyze the results of the questionnaire

We will mainly focus on the results of questions in the first part. We will attach a mark to each option (-2 for strongly disagree, -1 for disagree, 0 for neutral, 1 for agree, 2 for strongly agree). For

each question, we first calculate the total number of people who choose "I don't know". Then for other people who choose answers other than "I don't know", we will calculate the average mark of their answers. The average mark of answers to each question indicates whether what we built is useful and better than the current tools. The response rate places limitations on the effectiveness of the survey. A high response rate is crucial for generalizing the sample to the population. That's why we are offering inducement to increase the response rate.

5 Progress to date

1) We have done the architecture design and confirmed the technologies to use. The design style is to provide clear information and have fewer secondary selection buttons.
2) We have done the database schema design after several meetings with the client.
3) We have written insertion scripts for current data and imported real data into the database schema.
4) We have implemented some basic functions.
5) We have taken notes for all the client meetings and group meetings.
6) We have done the basic website UI design.
7) We have provided the basic function demo in proposal presentation.

6 Next steps

For next step, our team will focus on the function design and data analysis. We will prepare some specific questions based on the analysis, after that we can design some basic functions to answer those questions by generating the historical data. By the end of Semester 1, we hope we can accomplish some basic functions design and the overview of the website design including Data access and Business logic.
In general, we will finish the business logic and web service after confirming the requirements. We will have API service providing basic data. Business logic will handle these data and calculate different data after receiving the request from a browser or end user client. To ensure the safety of data, we will finish security authorization function for all web pages. In further, we may design responsive web application which can self-adapt to mobile devices. More critical questions will arise based on mining big data.
For group task allocation in Semester 1, Zeqi Fu will focus on the link between data and business logic and the link between business logic data and viewer layer which is the Controller layer in MVC architecture. Minh Tam Phan will keep eyes on further database design and business logic design which is related to Model layer in MVC architecture. Zeyu Lin will generate and design the website UI which is the View layer in MVC architecture.

7 Conclusion

This application is built based on the data collected from basic data sources of WebSubmission and Moodle systems. The research method survey research is used to evaluate whether what we built is useful and better than the current tools. This application can provide much help for the teaching staffs to improve the teaching practice based on the reports. Furthermore, this is a web application; hence, it is convenient to use.

A HEADINGS IN APPENDICES

A.1 Introduction and Motivation

A.2 Background

A.3 Architecture

A.4 Research Methodology

A.5 Progress to date

A.6 Next steps

A.7 Conclusion

A.8 References

REFERENCES

[1] Lonn, Steven, and Stephanie D. Teasley. "Saving time or innovating practice: Investigating perceptions and uses of Learning Management Systems." Computers & Education 53.3 (2009): 686-694.
[2] Álvarez, Pedro, et al. "Alignment of teacher's plan and students' use of LMS resources. Analysis of Moodle logs." Information Technology Based Higher Education and Training (ITHET), 2016 15th International Conference on. IEEE, 2016.
[3] Edwards, Stephen H., et al. "Comparing effective and ineffective behaviors of student programmers." Proceedings of the fifth international workshop on Computing education research workshop. ACM, 2009.
[4] Nandi, Dip, et al. "How active are students in online discussion forums?." Proceedings of the Thirteenth Australasian Computing Education Conference-Volume 114. Australian Computer Society, Inc., 2011.
[5] Ceddia, Jason, Judy Sheard, and Grant Tibbey. "WAT: a tool for classifying learning activities from a log file." Proceedings of the ninth Australasian conference on Computing education-Volume 66. Australian Computer Society, Inc., 2007.
[6] Costello, Eamon. "Opening up to open source: looking at how Moodle was adopted in higher education." Open Learning: The Journal of Open, Distance and e-Learning 28.3 (2013): 187-200.
[7] Krassa, Anna. "Gamified Moodle Course in a Corporate Environment." (2013): 84-93.
[8] Castro, Félix, et al. "Applying data mining techniques to e-learning problems." Evolution of teaching and learning paradigms in intelligent environment. Springer Berlin Heidelberg, 2007. 183-221.
[9] Romero, Cristóbal, Sebastián Ventura, and Enrique García. "Data mining in course management systems: Moodle case study and tutorial." Computers & Education 51.1 (2008): 368-384.
[10] Romero, Cristóbal, et al. "Data mining algorithms to classify students." Educational Data Mining 2008. 2008.
[11] Krüger, André, Agathe Merceron, and Benjamin Wolf. "A data model to ease analysis and mining of educational data." Educational Data Mining 2010. 2010.
[12] Zhang, Hangjin, and Kevin Almeroth. "Moodog: Tracking Student Activity in Online Course Management Systems." Journal of Interactive Learning Research 21.3 (2010): 407-429.
[13] Mazza, Riccardo, et al. "Moclog–monitoring online courses with log data." (2012): 132-139.
[14] Mazza, Riccardo, and Luca Botturi. "Monitoring an online course with the GISMO tool: A case study." Journal of Interactive Learning Research 18.2 (2007): 251.
[15] Dierenfeld, Helena, and Agathe Merceron. "Learning analytics with excel pivot tables." (2012): 115-121.
[16] Sampayo, F. C. "Analytics and Recommendations." Moodle Docs. Retrieved from https://moodle. org/plugins/view. php (2013).
[17] Konstantinidis, Andreas, and Cat Grafton. "Using excel macros to analyse moodle logs." (2013): 33-39.
[18] Casany Guerrero, María José, et al. "Analyzing moodle/lms logs to measure mobile access." UBICOMM 2012: The Sixth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies. 2012.

[19] Easterbrook, Steve, et al. "Selecting empirical methods for software engineering research." Guide to advanced empirical software engineering. Springer London, 2008. 285-311.
[20] Research Methods Knowledge Base: Types of Surveys. URL: https://www.socialresearchmethods.net/kb/survtype.php
[21] Research Methods Knowledge Base: Constructing the Survey, Types Of Questions. URL: https://www.socialresearchmethods.net/kb/questype.php