Global Journal of Engineering Education

Modelling students' activities in programming subjects through educational data mining

Mewati Ayub, Hapnes Toba, Steven Yong & Maresha C. Wijanto

Maranatha Christian University
Bandung, Indonesia

ABSTRACT: This research explores educational data mining (EDM) from a learning course management system (LMS) and academic system at the Faculty of Information Technology at Maranatha Christian University in Bandung, Indonesia. The main objective of this study was to discover whether the students have used the LMS effectively to complete their learning process and enhance academic achievements in their study. As case studies, this research combines data from two programming courses in an informatics Bachelor programme, which are: Introductory Programming (IP) and Algorithm and Data Structures (ADS). EDM techniques are used to extract interesting patterns in the form of association rules. Two sets of interesting rules for the IP course and three sets of rules for the ADS course are obtained. As final results, some suggestions are proposed to enhance the LMS. The main idea is to apply a gamification method in the blended-learning environment to increase motivation of the students to utilise their free time more effectively during study.

Keywords: Blended learning, course management system, educational data mining, association rules, gamification

INTRODUCTION

The usage of information technology in education, especially in e-learning systems, has been widely implemented in all over the world. Such systems offer great flexibilities to share and communicate between students and lecturers in courses. The Web-based learning process has produced a large amount of data that originated from student-lecturer interactions. To explore those data from the educational environment, educational data mining (EDM) has been commonly utilised. With EDM, one can obtain patterns and hidden knowledge in large collections of data to enhance the learning system quality [1][2].

Blended learning is a variant of e-learning, which combines face-to-face instruction with technology-mediated instruction [3]. Some benefits of blended learning are to increase learning effectiveness, convenience and access. This research explores EDM from a learning course management system (LMS) and academic system in the Faculty of Information Technology at Maranatha Christian University in Bandung, Indonesia. The faculty adopt a blended learning system with full face-to-face instruction. In this case, the LMS is used to complete the learning process. The objective of this study is to discover whether the students have used the LMS effectively to complete their learning process and enhance achievements in their study. As case studies, this research combines data from two programming courses in the informatics Bachelor programme, which are: Introductory Programming (IP) and Algorithm and Data Structures (ADS). The IP course is conducted in the first semester and the ADS course in the second semester. The students in the ADS course are a subset of the IP course. The findings from EDM of these data will be used as a foundation to improve the system.

EDUCATIONAL DATA MINING

Data mining has been applied to data from different types of educational systems. Data mining in education is also known as educational data mining. EDM is concerned with developing methods for exploring the unique types of data from educational environments. Those methods are used to improve understanding of students' behaviours and the system environments in which they are involved [4]. The educational system or the environment itself consists of traditional education and computer-based education or Web-based education. The traditional one is still the most widely used educational system [5]. Computer-based education is also called a learning management system (LMS), course management system (CMS) or learning content management system (LCMS) [2][6]. The differences between them are based on the data sources provided.

Those data sources need to be further processed depending on the nature of data and the problems that need to be resolved [5]. Romero and Ventura categorise the works in EDM into two main methods. The first is statistics and visualisation and the second is Web mining (including clustering, classification, association rule mining, sequential pattern mining, text mining and others) [5]. The Web mining methods are quite often implemented in EDM today. In addition, Baker classifies work in EDM into the following categories [4]:

1. Prediction:

    - Classification, regression and density estimation;

2. Clustering;
3. Relationship mining:

    - Association rules mining, correlation mining, sequential pattern mining and causal data mining;

4. Distillation of data for human judgement;
5. Discovery with models.

The first three categories are familiar to most researchers in data mining. EDM is an interdisciplinary area including, but not limited to, information retrieval, recommender system, social network analysis, and so on. In fact, EDM can be described as the combination of three main areas: computer science, education and statistics [5], which could be used as supporting tools in course and students' activity design [8].

Frequent patterns are patterns that appear frequently in a data set and are useful for discovering interesting relationships hidden in large data sets. Such relationships can be represented and uncovered in the form of association rules. Finding frequent patterns is important in mining associations, correlations, classification, clustering, and other data mining tasks. Support and confidence of rules are two measures of association rule. Those measures reflect the usefulness and certainty of the rules discovered. Support shows the frequency of item set that appears in dataset. Confidence shows the frequency of the rule is true or happens. For example, consider an association rule:

```
quiz = complete => grade = excellent [support =2%, confidence = 60%]
```

A support of 2% means that 2% of all the transactions show that students who have completed the quizzes and achieved an *excellent* grade in a course arose together. A confidence of 60% means that 60% of the students who have done the quizzes completely, have also passed the course successfully.

Association rules are considered to be reliable if they satisfy both a minimum support threshold and a minimum confidence threshold, and they are called strong rules. However, the support and confidence measures are insufficient for filtering out uninteresting association rules. To cover this, a correlation measure can be used. One of the simple correlation measures is what is known as lift.

The lift between the occurrence of A and B can be measured by the formulae in Equation (1).

```
lift(A, B) = P(A∪B) / P(A).P(B)
lift(A, B) = P(B|A) / P(B)                                    (1)
lift(A, B) = confidence(A= B) / support(B)
```

If the lift's value is less than 1; then, the occurrence of A is negatively correlated with the occurrence of B. If the lift's value is greater than 1; then, A and B are positively correlated. It means that the occurrence of A implies the occurrence of B. If the result is equal to 1, the A and B are independent, and there is no correlation between them [7].

METHODOLOGY

The study was based on data from two courses of programming subjects. The first course is Introductory Programming (IP) and the second is Algorithms and Data Structures I (ADS). Both courses are closely related, the IP course is a prerequisite of the ADS course. Students that follow the ADS course come from attendees of the IP course.

Students' activities data have been extracted from the LMS. Activity data attributes for each student are access time, Internet protocol address, user identity, action and activity information. As shown in Table 1, a group of attributes has been selected for EDM. These attributes consist of:

a) activity data from the LMS, such as user identity, access time, action and material;
b) academic data, such as course final grade and GPA of first semester;
c) session time, transformed from access time and course schedule;
d) activity level, transformed from activity frequency of a student.

Table 1: Students' activities data set.

Attribute name	Description	Possible values
Time	Access time	[Morning, afternoon, night]
Session	Session time	[In course, free time]
Action	Type of action	[resource view, doing exercise, quiz attempt]
Material	Material resources	Depend on the course
Activity	Activity level	[Low, medium, high]
Grade	Course final grade	[Excellent, good, fair, poor]
GPA	GPA of first semester	[Excellent, good, fair, poor]

The data set was built from 68 students from the IP course and 55 students from the ADS course. The students' activities data set was made up of 438 rows for the IP course and 990 rows for the ADS course. For the access time attribute, the morning session is from 5.00 am until 11.59 am, the afternoon is from 12.00 pm until 06.00 pm and others considered to be night. For the session time attribute, free time is the time outside the lectures. The activity level is counted as the frequency of students' activities in one semester in a course. This study explored students' activity data sets through association rules mining to obtain interesting rules. The rules can be utilised to improve the learning system and to encourage the students to be involved in more actively during the sessions. The experiment was conducted twice, one for each student's activities data set from the IP course and the other for data set from the ADS course.

RESULTS AND DISCUSSION

A histogram of students' access time of the IP course and the ADS course are shown in Figure 1. Most of the students' preferences access time are in the morning for both courses. In the IP course, more students have more preferences to access the system at night than in the afternoon, but *vice versa* in the ADS course. In Figure 2, the histogram of students' session time shows that most of the students prefer to access the LMS outside of the course, which is in their free time for both courses.

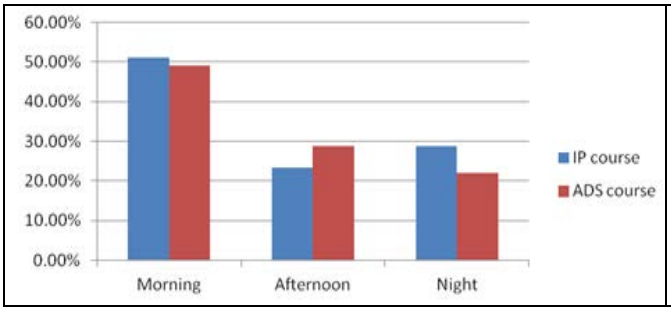

Figure 1: Histogram of access time.

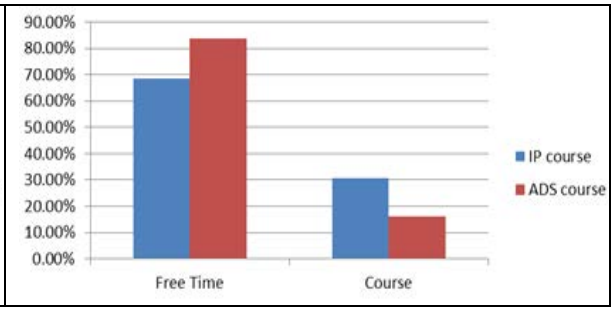

Figure 2: Histogram of session.

In Figure 3, the histogram of the students' final grade describes the grade distribution in each course. The IP course grade is dominated by fair, followed by good, excellent and poor. However, the excellent grade dominates the ADS course, followed by fair, good and poor. Students that follow the ADS course must pass the IP course. This condition might have caused the distinction of grade distribution between both courses.

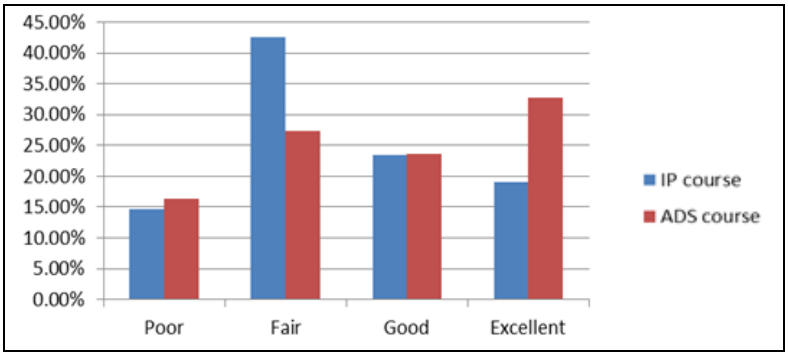

Figure 3: Histogram of course grade.

Using association rules mining against the students' activity data set, two sets of interesting rules for the IP course and three sets of rules for the ADS course were obtained. The data mining tools used during the experiments were WEKA version 3.8.1 and the Microsoft Azure Cloud Machine Learning System. To determine whether a rule is interesting, three parameters were considered: support, confidence and lift. The minimum support is set at 0.05, the minimum confidence was 0.75, and the lift had to be greater than 1.0. In Figure 4, an excerpt of the ADS data set is shown.

A set of rules about students' access time based on their grade in free time sessions is described in Table 2 for the IP course and in Table 3 for the ADS course. The preferred time to access learning resources of the IP course was at night for students that have fair, excellent and good grades. The alternative access time was in the afternoon. The preferred time to access learning resources of the ADS course was at night for those achieving a good grade, in the morning for those achieving an excellent grade and in the afternoon for those achieving a poor grade. Students, that access learning resources in free time, indicate that they have willingness to learn more outside the class.

	StudentID	Access Time	Session	Action	Material	Activity	GradeASD	GPA
2	1672051	Afternoo	Free Time	resource view	Slide List Linier	Medium	Good	Excellent
3	1672051	Morning	In course	resource view	Slide List Linier	Medium	Good	Excellent
4	1672051	Morning	Free Time	resource view	Slide ADT Stack	Medium	Good	Excellent
5	1672033	Afternoo	Free Time	resource view	Slide Sorting Lanjut	High	Fair	Fair
6	1672033	Morning	Free Time	resource view	Slide List Linier	High	Fair	Fair
7	1672033	Night	Free Time	resource view	Slide Stack Queue dengan List	High	Fair	Fair
8	1672033	Night	Free Time	resource view	Slide Variasi List	High	Fair	Fair
9	1672009	Afternoo	Free Time	resource view	Slide Sorting Lanjut	High	Fair	Good
10	1672009	Morning	In course	resource view	Slide Sorting Lanjut	High	Fair	Good

Figure 4: An excerpt of the data set.

	StudentID	Access Time	Session	Action	Material	Activity	GradeASD	GPA
2	1672023	Afternoon	Free Time	resource view	Slide Sorting Lanjut	Low	Excellent	Excellent
3	1672023	Afternoon	Free Time	resource view	Slide Sorting Lanjut	Low	Excellent	Excellent
4	1672066	Morning	Free Time	resource view	Slide List Linier	Low	Excellent	Excellent
5	1672066	Night	Free Time	resource view	Solusi Kuis2	Low	Excellent	Excellent
6	1672066	Afternoon	In course	resource view	Slide List Linier	Low	Excellent	Excellent
7	1672066	Morning	Free Time	resource view	Latihan Stack	Low	Excellent	Excellent
8	1672066	Morning	Free Time	resource view	Slide ADT Stack	Low	Excellent	Excellent
9	1672058	Night	Free Time	resource view	Solusi Kuis2	Low	Excellent	Excellent
10	1672058	Night	Free Time	resource view	Latihan Queue	Low	Excellent	Excellent

Figure 5: Some outliers of the activity data set.

Table 2: Rules of students' access time of the IP course based on GradeIP.

Rule #	Rule	Parameters		
		Support	Confidence	Lift
1	GradeIP = Fair, Time = Night => Session = Free Time	0.137	1.000	1.446
2	GradeIP = Fair, Time = Afternoon => Session = Free Time	0.087	0.93	1.35
3	GradeIP = Excellent, Time = Night => Session = Free Time	0.059	1.000	1.446
4	GradeIP = Excellent, Time = Afternoon => Session = Free Time	0.059	0.963	1.392
5	GradeIP = Good, Time = Night => Session = Free Time	0.050	1.000	1.446

Table 3: Rule of students' access time of the ADS course based on GradeADS.

Rule #	Rule	Parameters		
		Support	Confidence	Lift
1	GradeADS = Good, Time = Night => Session = Free Time	0.118	1.000	1.19
4	GradeADS = Excellent, Time = Morning=> Session = Free Time	0.098	0.87	1.03
5	GradeADS = Poor, Time = Afternoon => Session = Free Time	0.057	0.95	1.13

Activity level is the frequency of student access to learning resources. Having high activity in free time sessions indicates willingness to learn more outside the class. A set of rules about students' activity based on their grade in free time session is shown in Table 4 for the IP course and in Table 5 for the ADS course.

Table 4: Rule of students' activity of the IP course based on GradeIP.

Rule #	Rule	Parameters		
		Support	Confidence	Lift
1	GradeIP = Fair, Activity = High => Session = Free Time	0.240	0.761	1.100
2	GradeIP = Excellent, Activity = Medium => Session = Free Time	0.096	0.933	1.349
3	GradeIP = Excellent, Activity = High => Session = Free Time	0.059	1.000	1.446

In the IP course, students that had high activity, achieved fair and excellent grades. Some excellent students also had medium activity. In the ADS course, students that had high activity, achieved good, fair and excellent grades, while students that had medium activity, achieved fair, good and poor grades.

Table 5: Rule of students' activity of the ADS course based on GradeADS.

Rule #	Rule	Parameters		
		Support	Confidence	Lift
1	GradeADS = Good, Activity = High => Session = Free Time	0.168	0.85	1.02
2	GradeADS = Fair, Activity = High => Session = Free Time	0.129	0.89	1.06
3	GradeADS = Fair, Activity = Medium => Session = Free Time	0.124	0.87	1.04
4	GradeADS = Good, Activity = Medium => Session = Free Time	0.106	0.88	1.05
5	GradeADS = Excellent, Activity = High => Session = Free Time	0.078	0.87	1.03
6	GradeADS = Poor, Activity = Medium => Session = Free Time	0.057	0.95	1.13

Table 6 and Table 7 describe a set of rules about students' access time and activity of the ADS course based on their GPA. The preferred time for accessing learning resources for good GPA was in the morning and at night. Some of fair GPA students preferred to access the LMS in the afternoon, while excellent GPA students accessed at night. In Table 7, only good GPA students had high activity, while medium activity was noted by good and fair GPA students.

Although there is an assumption that students who have an excellent grade will also have high activity in learning, the activity data set shows some outliers. In Figure 5, there were some students who achieved an excellent grade and GPA, but had low activity. This does not mean that the students were not willing to use the LMS. It is likely that they have downloaded some study materials only once or copied them from their classmate and, then, they studied the material at their favoured time, outside the system.

Table 6: Rule of students' access time of the ADS course based on GPA.

Rule #	Rule	Parameters		
		Support	Confidence	Lift
1	GPA = Good, Time = Morning => Session = Free Time	0.160	0.91	1.09
2	GPA = Good, Time = Night => Session = Free Time	0.110	1.00	1.19
3	GPA = Fair, Time = Afternoon => Session = Free Time	0.090	0.92	1.09
4	GPA = Excellent, Time = Night => Session = Free Time	0.083	1.00	1.19

Table 7: Rule of students' activity of the ADS course based on GPA.

Rule #	Rule	Parameters		
		Support	Confidence	Lift
1	GPA = Good, Activity = Medium => Session = Free Time	0.174	0.91	1.09
2	GPA = Good, Activity = High => Session = Free Time	0.168	0.93	1.11
3	GPA = Fair, Activity = Medium => Session = Free Time	0.063	0.90	1.07

Some suggestions can be proposed to enhance the learning system. The main idea is to apply a gamification method in the blended-learning environment. The gamification method is expected to be able to increase motivation of the students so that they to utilise their free time more effective in study. The suggested main features are given in Table 8.

Table 8: Suggested main features of the gamification method.

Rating	Every student gives a rating about material/resources, such as difficulty level and fills in a questionnaire of a course.
Achievement	Achievements, which have been reached by the students based on their activities. These are shown using colour, emoticon, certificate and sound chat.
Notification	Notifications are given during login, such as about quiz, pre-test, post-test, new resources.
Event	To give variations in activities for the students, such as games about the study materials, create drag and drop quizzes.
Leaderboard	To show prestige of some students, such as top ten students based on their achievement.
Tournament	To organise a competition and provide some additional challenges in the learning environment. Tournament could be used to measure students' involvement, groups' creativity and engagements to each other.
Forecasting	To forecast the final grade based on students' activities and achievements. This feature is shown after the mid semester examination. The feature can be used to give suggestions about the resources that must be learned more intensive to reach a better grade.

CONCLUSIONS

In this research, EDM techniques have been explored to find interesting rules in a student activities data set. The study reveals that there are strong correlations between students' access time, their activities in the LMS and their final grade. It is important to create an LMS which could attract students' enthusiasm in taking extra efforts for the success of their

study. A gamification method in the blended-learning environment is suggested as the final finding of this research. The gamification method is expected to be able to increase motivation of the students to utilise their free time more effectively during study.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support provided by the Directorate General of Research and Development Strengthening in the Ministry of Research, Technology and Higher Education of the Republic of Indonesia, under the Research Grant number 1598/K4/KM/2017.

REFERENCES

1. Romero, C. and Ventura, S., Educational data mining: a survey from 1995 to 2005. *Expert System with Applications*, 33, **1**, 135-146 (2007).
2. Romero, C., Ventura, S. and Garcia, E., Data mining in course management systems: Moodle case study and tutorial. *Computers and Educ.*, 51, **1**, 368–384 (2008).
3. Graham, C.R., *Blended Learning Models*. In: Encyclopedia of Information Science and Technology. Hershey, PA: Idea Group Inc., 375-383 (2009).
4. Baker, R. and Yacef, K., The state of educational data mining in 2009: a review and future visions. *J. of Educational Data Mining*, 1, **1**, 3-16 (2009).
5. Romero, C. and Ventura, S., Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3, **1**, 12-27 (2013).
6. Pankin, J., Roberts, J. and Savio, M., *Blended learning at MIT*. Massachusetts Institute of Technology Repository. (2015).
7. Han, J., Kamber, M. and Pei, J., *Data Mining Concepts and Techniques.* Waltham: Elsevier, Inc., 264-266 (2012).
8. Stewart, M.F. and Chisholm, C.U., Comparative analysis of emotional competency within distinct student cohorts. *Global J. of Engng. Educ*., 14, **2**, 163-169 (2012).

BIOGRAPHIES

Mewati Ayub graduated with a Bachelor of Informatics from Bandung Institute of Technology (ITB) in 1986. She completed her Master's degree at Bandung Institute of Technology in 1996 and her doctoral degree at Bandung Institute of Technology in 2006. She has been working as a faculty member in the Faculty of Information Technology at Maranatha Christian University, Bandung, Indonesia, since 2006. Her specialty is in the field of educational technology, software engineering and data mining.

Hapnes Toba graduated with a Master of Science from Delft University of Technology, the Netherlands in 2002. He completed his doctoral degree at Universitas Indonesia in 2015. He has been working as a faculty member in the Faculty of Information Technology at Maranatha Christian University, Bandung, Indonesia, since 2003. His specialty is in the fields of information retrieval, text processing and data mining.

Steven Yong has been a student of the informatics study programme of the Faculty of Information Technology, Maranatha Christian University, Bandung, Indonesia, since 2014. Now he is working on his final-year project.

Maresha Caroline Wijanto is an alumnus of the Faculty of Information Technology, Maranatha Christian University and also graduated from Bandung Institute of Technology (ITB) with her Master's degree in computer science. She has been working as a faculty member in the Faculty of Information Technology at Maranatha Christian University, Bandung, Indonesia, since 2010. Her specialty is in the field of natural language processing and data mining.