GROUPING STUDENTS' LEARNING PATTERNS WITH MANABA'S LOG DATA BY K-MEANS

Kai Li
Dokkyo University
1-1 gukuen-cho, Soka, Japan

ABSTRACT

Assessing students' performance in online learning could be executed not only by the traditional forms of summative assessments such as using essays, assignments, and a final exam, etc. but also by more formative assessment approaches such as interaction activities, forum posts, etc. However, it is difficult for teachers to monitor and assess students' learning activities using the log data. To provide teachers with a more comprehensive view of students' distinct learning behaviour patterns, and to supply personalized interventions and support to meet the specific needs of each learning group, this study focuses on how to automatically acquire learning logs from Manaba, a Japanese commercial LMS, and how to cluster students' learning activities using the k-means algorithm. Firstly, we developed a program using Python to scrape students' learning activity log information from the Manaba web pages. We collected 56446 lines of clickstreams log data from 121 students in two computer literacy hybrid classes in the fall semester of 2022 (2022/9~2023/1). Secondly, we convert the raw logs into a structured dataset with 33 features which represent each student's learning activities. Then we extract and select 15 features representing three perspectives: raw activity, time on task, and learning frequency. Thirdly, we grouped students' learning activity patterns with the three perspectives into 5 clusters by the k-means clustering algorithm. As a result, this study identified five distinct learning activity patterns depending on how much, how long and how often the students learned online. For example, cluster 1 seldom learned but spent time on learning whom we considered the disengaged or struggling students, and cluster 5 had more learning activities with little time on each activity whom we considered the well-self-regulated students. The results of this study contribute to how to monitor students' learning activity in online learning and how to assess and support student's learning by their learning activity patterns.

KEYWORDS

Manaba, Learning Management System, Learning Analytics, Learning Activity, Clustering, K-Means

1. INTRODUCTION

Due to the COVID-19 pandemic, all teaching has moved from a face-to-face environment to a fully remote environment for educational institutions worldwide (Bradley, 2021). According to the United Nations (2020), ―by mid-April 2020, 94 per cent of learners worldwide were affected by the pandemic, representing 1.58 billion children and youth, from pre-primary to higher education, in 200 countries. On 7th April 2020, the Government of Japan declared a state of emergency concerning COVID-19, education institutions were faced with decisions about how to continue teaching and learning while keeping their faculty, staff, and students safe. Most institutions have opted to cancel all face-to-face classes and have mandated that faculty move their courses online to help prevent the spread of the virus. Evidence has demonstrated that universities took to online environments in a bid to save the 2020 academic year through the use of the Internet and digital platforms (Bao, 2020; Crawford et al., 2020). Remote learning or e-learning is more important than ever due to the increased home study in many countries prompted by the COVID-19 pandemic (Paudel, 2021). Other academic studies focused on distance education activities (Aristovnik et al., 2020), teacher and student experiences (Meda & ElSayary, 2021), success, perception, and attitude regarding the online learning process (Wang et al., 2020).

Learning Management System (LMS) could be used for course management for example to publish learning content, quizzes, attendance, forums, etc. LMS could also be used for collecting learning activity data for example history of accessed pages, quiz scores, and assignment submission date (Romero et al.,

2008). Students' learning experience in an LMS allows the teacher to gather feedback from students and to monitor data such as the amount of time spent online, and which pages were accessed. The learner's learning activity data accumulated in the LMS is known as learning log data. Well-designed learning activities can be used to monitor student progress and provide data that helps teachers identify each student's strengths and weaknesses (Yassine et al., 2016). There has been increasing research that analyses learning logs to improve educational effects (Aldowah et al., 2019, Bachhal et al., 2021).

Learning analytics is a form of data analysis that allows teachers to look for students' online traces and information associated with the learning processes. The major area of applying learning analytics is the act of predicting and monitoring learning performance (Bichsel, 2012), and offers feedback to prevent poor performance and eventual failure of students (Pardo et al., 2017). Some researchers also tried to detect students' learning behaviour in LMS to create personalized learning to fit the characteristics of students in achieving better learning outcomes (Purwoningsih et al., 2019).

To understand students' online learning activities, the K-means clustering algorithm is a simple and effective way to group students with similar behaviours. Some researchers used K-means to cluster patterns of students' behaviours in an interactive online mathematics game, and the results indicated that students in four clusters, except for slow progressors, showed significant increases in their understanding of mathematical equivalence (Lee et al., 2022). Other researchers discovered groups of students enrolled in the emergency remote teaching online course based on the various course-related data collected throughout the first year of the COVID-19 pandemic using K-means and identified distinct groups of students for future adaptations of the online course design to improve the retention and their final grades (Balaban et al., 2023). In addition, some research applied K-means to explore learner profiles in terms of how they performed self-regulated learning (SRL) in Massive Open Online Courses (MOOCs). They revealed four different self-regulated learner profiles and identified cultural differences between those clusters (Tang, 2021).

Manaba is a Japanese commercial cloud-based LMS developed by AsahiNET, Inc. in 2007. It is reported that 250 educational institute users use Manaba by September 2020. Since the access and storage limitation of the former LMSs, Dokkyo University newly contracted Manaba as the main LMS after the spread of COVID-19 in May 2020. As the other LMS, Manaba could be used for course management to publish learning contents, quizzes, attendance, forums, etc., and could be used for collecting learning activity data for example history of accessed pages, quiz scores, and assignment submit date. However, Manaba could only display log data page by page, but not download it, which has no meaning for analysing students' learning activities in detail.

To provide teachers with a more comprehensive view of students' distinct learning behaviour patterns in Manaba, and to supply appropriate interventions and personalized support to meet the specific needs of each learning group, we generated the following research questions:

1. How to collect the log data from Manaba? How to automate the data collection to reduce the workload for teachers while enabling them to further teaching and learning support.
2. How to group students on their online learning activities with the raw log data? How to remote monitor students' learning activity to identify early signs of disengagement or intervene appropriately to support struggling students.

2. METHODOLOGY

2.1 Data Collection

Firstly, we developed an RPA (Robotic Process Automation) program using Python to collect the log data from Manaba. The program is developed with the Selenium library and the Chrome Driver with Python. Selenium library could be used to achieve the same movements as humans operate via a browser. Since each tag, link, and element is defined in a variety of ways on the web pages, depending on the definition of each element, we used find_element_by_id, name, tag, link, and XPath methods with the Selenium library to search and collect each log data that showed on Manaba. Since scraping information on the Internet is limited to "private use" and "information analysis" by law, source code is not opened in this study. It could be requested if needed.

With the program, it could automatically login to the Manaba, select the target course, transit to the target pages, search log data on the pages, move to the next pages and record all the log data into a CSV file. We collected 56446 lines of clickstreams log data from 121 students in two computer literacy hybrid classes in the fall semester of 2022 (2022/9~2023/1). Each data included the information of the student's name, ID, access time, URL, function category of the page, page type, and the page's title. The "Function category of the page" includes 10 functions like course top, report, contents, forum, mini-test, grades, etc. The "Type of the page" includes 17 types like course top, submit, input upload, start page, top page, etc.

In the previous research, we visualized and analysed students' clickstreams and page transit trajectories with the Retentioneering Python library. By retentionneering, we could visualize which learning pages were connected by students' transitions. And by the weight of the edges between each page, we could understand how many students transit from one learning page to the other learning pages. We also used a step matrix to show the sequential learning pages that the students accessed step by step. By the step matrix, we could know how many steps had been passed before the remote learning ended, and whether all the necessary learning pages were accessed before the learning ended. We also clustered students' learning behaviours into 6 clusters by their page transition patterns. The cluster divided students' main learning patterns into content-centred and report-centred, comment or not. However, we do not know how long the students spend on each learning page by the Retentioneering library. In this study, we will focus on the learning time on each page and cluster their learning patterns.

2.2 Data Preprocessing

Secondly, we calculated the time spent on each page and transformed the raw activity log data into a structured dataset, then we identified the important variables in three perspectives by removing the redundant ones. Finally, we clustered students' learning behaviour by the k-means cluster algorithm. The data manipulation and analysis procedures were implemented using Python and Scikit-learn.

2.2.1 Structurization and Standardization

Depending on Santos's research (2023), they extracted 30 variables from three perspectives using Moodle's log data. In this study, we transform the Manaba log data into a structured dataset with 33 variables in the raw activity, time on task, and frequency perspectives. The raw activity includes all the variables that we considered to count the number of times a certain action is performed by the student on Manaba. Time on task stores all variables associated with the amount of time spent on each activity and frequency stores all variables that, to some degree measure how often students access Manaba. By the log data of the function category of the page and the type of the page, we select 22 variables in the raw activity perspective including the click count of Total Click, CourseTop, Collection, Contents, Project, Report, Forum, MiniTest, Scores, questionnaire, List of Comments, Submit Comments, List of Thread, Top Page, Page, List of Assignments, Start Page, Submit Report, Submit Files, Submit Cancel, Attachment Files, and Input Upload. The time on task perspective got 7 variables including the stay time spent on Questionnaire, CourseTop, Contents, Report, Forum, Total Time Over1200sec, and Largest period of inactivity (Day). The frequency perspective got 4 variables including Clicks/day, Days with > 10 clicks, Days with 0 clicks, and Days with 0 clicks (% of period). The time spent on each activity over 20 minutes (1200 seconds) was excluded from the data, since an activity leaving the computer over 20 minutes may not be a learning activity, but just leaving the computer aside. The resulting transformed structured data frame had 121 rows and 33 columns, with each row representing a student and each column representing a learning activity variable.

Due to the differences in scale in each variable, we normalized all the data to standardised scores (mean =0 and standard deviation =1). It's important to scale the variables to ensure that they are on the same scale. This also helps the k-means algorithm to work effectively.

2.2.2 Variable Selection

Variable or feature selection is the process of reducing the number of input variables in machine learning. It is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model (Kuhn, 2013). Of the 33 variables, some are relevant, and some have a low contribution to the results, we need to remove the variables with a low correlation and low explained variance.

First, to identify the relevant variables, we used the Spearman correlation index and set an absolute value threshold (0.7) as the cut-off for relevant variables, then the remaining variables could be removed. Second, to identify the low contribute variables, we apply K-Means clustering separately on each perspective to create perspective-specific clusters. And we calculate the explained variance for each perspective. The explained variance measures how much variance in the data is explained by each cluster. This can help to identify which clusters are more meaningful and distinct. For each perspective, we identified the variables associated with clusters that have the lowest explained variance. Variables that have low variance across clusters may not be informative for distinguishing between groups. Based on the results, we removed the variables associated with clusters that have the lowest explained variance. As a result, we select 15 variables for the final clustering (Table 1). Each row gives the centroid coordinate for each cluster which means the average value for a variable in the cluster.

Table 1. K-means standardized mean for all variables used in clustering

Perspective	**Variable**	**cluster 1** n=125(21%)	**cluster 2** n=3(2%)	**cluster 3** n=57(47%)	**cluster 4** n=23(19%)	**cluster 5** n=13(11%)
Raw Activity (click count)	Course Top	-0.319	0.3599	0.0617	-0.7573	1.6208
	Contents	-0.2267	0.5164	0.0931	-1.0326	1.7443
	Forum	-0.4169	-0.3893	-0.0914	0.181	0.9914
	Mini Test	-0.3399	3.9931	-0.3238	-0.2968	0.5134
	questionnaire	-0.5817	-0.229	-0.5944	-0.4574	-0.1713
	List of Comments	-0.3223	-0.2532	0.0083	-0.0372	0.7104
	Submits of Comments	-0.9585	-0.9585	-0.8604	-0.6668	-0.3564
	Page	-0.2889	0.2896	0.0793	-0.9404	1.8057
	Submit Report	-0.1646	0.3417	0.404	-1.0085	-0.0016
	Submit Cancel	-0.9281	-0.287	-0.2373	-0.6969	-0.7947
Time on task (stay time)	Questionnaire	-0.1138	-0.0538	-0.067	0.3213	-0.0105
	Contents	0.1988	0.603	-0.4729	1.1732	-0.486
	Report	1.8478	-0.5536	-0.5203	-0.3099	-0.5536
	Forum	0.0295	-0.3787	-0.0578	0.0833	0.1521
Frequency	Days with 0 clicks	0.2962	-0.4396	-0.064	0.7172	-1.4948

2.3 Clustering

The k-means clustering algorithm is a simple and the most popular data mining technique on unsupervised data sets. It aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid). We used k-means to segment students into different groups based on their learning behaviours. This can help identify distinct learning patterns among different student groups for later assessing and improving students' learning performance. In the clustering process, the most important way is to identify the number of clusters (k) at the beginning. If the number of initial clusters specified is not good, then the clusters' results are not as they should be. The number of K was set to 5 with the assistance of the elbow method (Humaira and Rasyidah, 2020). In total, we have 21% of students in Cluster 1, 2% in Cluster 2, 47% in Cluster 3, 19% in Cluster 4, and 11% in Cluster 5 (Table 1).

Table 2. Means of each perspective in five clusters

Perspective	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**
Raw Activity	-0.43173	0.416135	-0.20345	-0.50093	0.857307
Time on task	0.144908	-0.05474	-0.15855	-0.07292	-0.28239
Frequency	0.296241	-0.43956	-0.06396	0.717188	-1.49481

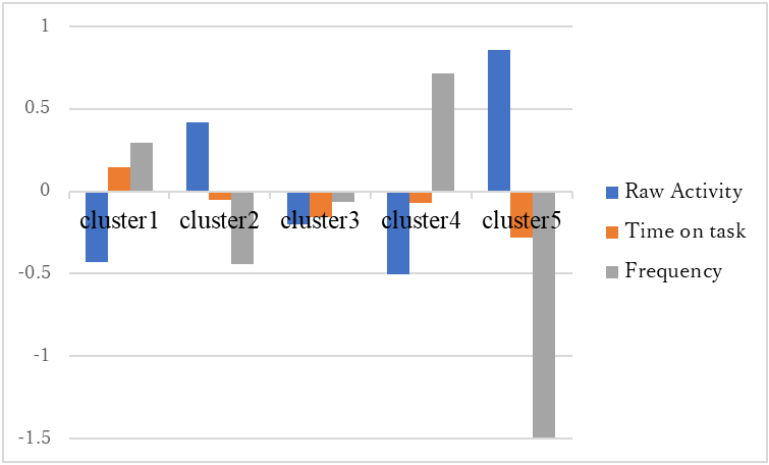

Figure 1. Perspective means in each cluster

Table 2 showcases the standardized group mean value for each of the variables used in each cluster. To better interpret our results, we described each student group according to their three distinctive perspectives to understand how students are grouped based on their activity patterns (Figure 1). Since the variable in the frequency perspectives shows the days with 0 clicks, the inverted data means the days with learning activities by which we concerned.

Depending on the characteristics of the three perspectives in each cluster, we named Cluster 1 as seldom learning but spending time on little activities group (little raw activity, more time on task, and more frequency with 0 click days). Cluster 2 was named always learning with more activities group (more raw activity, little 0 click days). Cluster 3 little activity and little time group (little raw activity and little time on task). Cluster 4 was named seldom learning and spending little time on little activities group (little raw activity, little time on task, and high frequency with 0 click days). Cluster 5 was named more learning with more activities but with little time on each activity (more raw activity, little time on task, and little frequency with 0 click days).

From a general point of view, cluster 5 is the self-regulation learners who learn much more often, and quicker with more activities. Cluster 2 is the common learners who learn with some activities. Cluster 3 is the next common learner with little activity and little time. Cluster 1 and Cluster 4 are similar, they seldom learn with little activities. Cluster 1 spent time on little activities, and Cluster 4 did not spend time on little activities. From the clusters, we found that students' remote learning behaviours were complex. Each group has different learning patterns which will lead to different assessments and support.

Table 3. Post Hoc comparisons between clusters

Perspective			**t**	**p**
Raw Activity	Cluster1	Cluster2	-3.486	0.006**
		Cluster5	-9.470	< .001***
	Cluster2	Cluster4	3.753	0.003**
	Cluster3	Cluster4	3.025	0.025*
		Cluster5	-8.670	< .001***
	Cluster4	Cluster5	-9.834	< .001***
Time on task	Cluster1	Cluster3	2.842	0.041*
		Cluster5	2.808	0.045*
Frequency	Cluster1	Cluster5	6.393	< .001***
	Cluster3	Cluster4	-3.860	0.002**
		Cluster5	5.682	< .001***
	Cluster4	Cluster5	7.781	< .001***

* p < .05, ** p < .01, *** p < .001

2.4 Comparison

To compare the differences in each perspective between clusters, we also conducted an ANOVA(Analysis of Variance) and post hoc comparisons(Table 3). For the raw activity perspective, cluster 5 >2 > 3 > 1> 4. For the spending time on each task, cluster 1 > 2 > 4 > 3 >5, and for the learning frequency (the frequency with 0 click days is inverted) 5 > 2 > 3 >1 > 4. Cluster 5 has more learning activities than the other clusters. Cluster 1 spent more time on learning tasks, and Cluster 5 learned frequently. The significant differences between each cluster can be confirmed in Table 3.

3. DISCUSSION

This study grouped students' learning activity patterns with the three perspectives into 5 clusters by the k-means clustering algorithm. We found that cluster 5 who are self-regulation learners learnt much more often, and quicker with more activities. The raw activity perspective is more than other clusters. They have more access to learning content, pages, tests etc. and have more interaction on the forum and submitting comments (Table 1). However, students of cluster 5 spent little time on learning content and they learnt more often than other clusters. By which we could assume that learning content is easy for them, the content is not enough, and other further reading materials should be supplied to these students. Another finding is cluster 1 and cluster 4 have few learning activities and seldom learn. Cluster 1 spent time on learning which we could assume that the learning content is difficult for them. They need personalized learning support. Cluster 4 spent little time and seldom learned who may drop out from learning. Teachers need to intervene to improve their learning engagement. In addition, cluster 1 (21%) and cluster 4 (19%) are made up of 40% of the students which shows the students have some struggles in online learning. Especially during the COVID-19 emergency, most of the students have mental issues and have little chance to communicate with others, therefore teachers should supply more feedback and support to sustain their online learning. We also found all the students seldom submit comments. Although the students of cluster 5 have more access to the list of comments, they just like to confirm others' comments but not submit comments. To support communication between students and teachers in online learning, teachers should supply timely personalized feedback in different ways. Overall, from the different learning activity patterns, teachers should assess students' performance not only by summative assessments such as using essays, assignments, and a final exam but also by more formative assessment approaches such as interaction activities, forum posts, etc. through their learning process.

4. CONCLUSION

The purpose of this study is to monitor students' remote learning process and assess their learning performance by their learning activity patterns with the learning log in Manaba. We developed a program which could automatically acquire learning logs from Manaba, and we clustered students' learning activities into five groups using the k-means algorithm. By clustering students' learning patterns, teachers could assess students' performance and optimize their teaching pedagogies. This study is based on Manaba's log, it could be adapted and replicated in other LMSs. By analysing students' learning activity patterns, educators and instructional designers can make data-driven decisions to enhance the design and delivery of online courses. This study contributes to improvements in content organization, interactivity, and engagement strategies. Also monitoring students' learning activities could help educators identify early signs of disengagement or challenges and intervene proactively to support struggling students. However, this study has some limitations. First, the automation program we developed with the Selenium library was slow. The execution time usually took 30 minutes to get 60 students' log data page by page from one course. The program needs to be improved with other faster libraries in the future. Second, the sample size is small which includes just 121 students in two classes, and the log data is only for one year's learning activities which need long-term research to get more learning activities. Third, we do not know the relationship between learning patterns and learning outcome performance. Further research should be analysed to predict students' learning performances and detection of undesirable learning behaviours with the log data. For example, using labelled

data with outcomes (e.g., student performance, and course completion) to build predictive models by machine learning algorithms to predict students' performances based on their learning patterns. Furthermore, how to intervene or support each group is not discussed in the study. Since LMS's functions are different and pedagogy is different, intervention methods should be considered in different environments.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 21K00655.

REFERENCES

Aldowah, H. et al, 2019. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, 37, pp. 13–49

Aristovnik, A. et al, 2020. A bibliometric analysis of Covid-19 across science and social science research landscape. *Sustainability*, 12(21), 9132. https://doi.org/10.3390/su12219132

Bachhal, P. et al, 2021. Educational data mining: A review. *Journal of Physics: Conference Series*, 1950(1)

Balaban, I. et al, 2023. Post hoc identification of student groups: Combining user modeling with cluster analysis. *Educ Inf Technol* 28, pp. 7265–7290. https://doi.org/10.1007/s10639-022-11468-9

Bao, W. (2020). COVID-19 and online teaching in higher education: A case study of Peking University. *Human Behavior and Emerging Technologies, 2*(2), pp. 113-115. https://doi.org/10.1002/hbe2.191

Bichsel, J. 2012. Analytics in Higher Education: Benefits, Barriers, Progress, and Recommendations (Research Report). *Louisville, CO Educ. Cent. Appl. Res*

Bradley, V. M. 2021. Learning management system (LMS) use with online instruction. *International Journal of Technology in Education*, 4(1), pp. 68–92

ChromeDriver, Retrieved from https://sites.google.com/chromium.org/driver/, last accessed 12/8/2023

Crawford, J. et al, 2020. COVID-19: 20 countries' higher education intra-period digital pedagogy responses. *Journal of Applied Learning & Teaching, 3*(1), pp. 1-20. https://doi.org/10.37074/jalt.2020.3.1.7

Humaira, H., & Rasyidah, R., 2018. Determining the appropriate cluster number using Elbow method for K-Means algorithm. *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA)* 2018, pp. 24-25

Kuhn, M., & Johnson, K., 2013. *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.

Lee, J. E. et al, 2022. Does slow and steady win the race?: Clustering patterns of students' behaviors in an interactive online mathematics game. *Educational technology research and development*, 70(5), pp. 1575-1599.

Manaba, Retrieved from https://ja.wikipedia.org/wiki/Manaba, last accessed 12/8/2023

Meda, L., & ElSayary, A., 2021. Establishing Social, Cognitive and Teacher Presences during Emergency Remote Teaching: Reflections of Certified Online Instructors in the United Arab Emirates. *Contemporary Educational Technology*, 13(4).

Pardo, A. et al, 2017. Provision of Data-Driven Student Feedback in LA & EDM, *Handbook of Learning Analytics*, pp. 163–174

Paudel, P., 2021. Online education: Benefits, challenges, and strategies during and after COVID-19 in higher education. I*nternational Journal on Studies in Education*, 3(2), pp. 70–85

Purwoningsih, T. et al, 2019. Online Learners' Behaviors Detection Using Exploratory Data Analysis and Machine Learning Approach, *Proceedings of 2019 Fourth International Conference on Informatics and Computing (ICIC)*, pp. 1-8

Retentioneering Library, Retrieved from https://github.com/retentioneering/retentioneering-tools, last accessed 12/8/2023

Romero, C. et al, 2008. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1), pp. 368–384

Santos, R., & Henriques, R., 2023. Grouping Bachelor's Students According To Their Moodle Interaction Profiles: A K-Means Clustering Approach. In L. Gómez Chova, C. González Martínez, & J. Lees (Eds.), 15th *International Conference on Education and New Learning Technologies*, pp. 7383-7389

Selenium Library, Retrieved from https://pypi.org/project/selenium/, last accessed 12/8/2023

Tang, H., 2021.Person-centred analysis of self-regulated learner profiles in MOOCs: a cultural perspective. *Education Tech Research Dev* 69, pp. 1247–1269. https://doi.org/10.1007/s11423-021-09939-w

United Nations, 2020. *Policy brief: Education during COVID-19 and beyond*. Retrieved from https://www.un.org/development/desa/dspd/wp-content/uploads/sites/22/20 Post hoc identification of student groups: Combining user modelling with cluster analysis 20/08/sg_policy_brief_covid-19_and_education_august_2020.pdf, last accessed 21/9/2023

Wang, C et al, 2020. Association between medical students' prior experiences and perceptions of formal online education developed in response to COVID-19: A cross-sectional study in China. *BMJ open*, 10(10), e041886.

Yassine, S. et al, 2016. A framework for learning analytics in Moodle for assessing course outcomes. *Proceedings IEEE Global Engineering Education Conference (educon)*, pp. 261-266.