Educational Technology International
2017, Vol. 18, No. 2, 159-191

Comparisons on Clustering Methods: Use of LMS Log Variables on Academic Courses*

Il-Hyun Jo
Ewha Womans Univ.

Yeonjeong PARK**
Honam Univ.

Jongwoo SONG
Ewha Womans Univ.

Korea

Academic analytics guides university decision-makers to assign limited resources more effectively. Especially, diverse academic courses clustered by the usage patterns and levels on Learning Management System(LMS) help understanding instructors' pedagogical approach and the integration level of technologies. Further, the clustering results can contribute deciding proper range and levels of financial and technical supports. However, in spite of diverse analytic methodologies, clustering analysis methods often provide different results. The purpose of this study is to present implications by using three different clustering analysis including Gaussian Mixture Model, K-Means clustering, and Hierarchical clustering. As a case, we have clustered academic courses based on the usage levels and patterns of LMS in higher education using those three clustering techniques. In this study, 2,639 courses opened during 2013 fall semester in a large private university located in South Korea were analyzed with 13 observation variables that represent the characteristics of academic courses. The results of analysis show that the strengths and weakness of each clustering analysis and suggest that academic leaders and university staff should look into the usage levels and patterns of LMS with more elaborated view and take an integrated approach with different analytic methods for their strategic decision on development of LMS.

Keywords : Learning management system, Academic analytics, Clustering analysis, Gaussian mixture models, K-means clustering, Hierarchical clustering

---

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A5B6036244).

** Corresponding author, ypark@honam.ac.kr

Introduction

Currently, Learning Management System (LMS) is getting ubiquitous in higher education (Krumm, Waddington, Teasley, & Lonn, 2014). Whether focusing on campus-based learning in higher institute for large cohorts of learners or distance learning from one-on-one tutoring to MOOC(Massive Open Online Course) environment, LMS is considered as an essential technology for virtual learning environment on e-learning systems where instructors or tutors provide various learning materials such as text, images, URL links and video clips to learners. Also the interaction between learners and system allows an adapted and personalized learning experience (Brooks, Greer, & Gutwin, 2014).

A common goal of LMSs is to organize and manage different courses within an integrated system. The integrated systems collect each learner's online behavior data in every class. Based on this data, educational researchers and practitioners are able to analyze and interpret students' learning patterns and progress during the semester. University staff and decision makers can leverage such LMS usage trends analytics to derive proper treatment and policies to their students.

Such a data-driven or data-assisted approach has been attempted in the field of higher education recently with the term of *academic analytics*. It has emerged after the widespread of data mining practices by the influence of business intelligence (Baepler & Murdoch, 2010; Goldstein & Katz, 2005). This approach has been evaluated as a new tool to respond to increased concerns for accountability in higher education and to develop actionable intelligence to improve student success and learning environment (Campbell, DeBlois, & Oblinger, 2007). For example, instructors and academic consultants are better able to understand the learner's learning behavior and performance, even their thoughts based on the rich data. Further, the academic analytics can help more strategic investment and development in a way to fulfill the needs of students and instructors based on the

informed analytic results via the pattern-recognition, classification, and prediction algorithms (Arnold, 2010).

The data analytics in education has helped to develop prediction models for academic success of learners based on their behaviors and participation or identifying at-risk students for special guidance from their faculty and advisors (Arnold & Pistilli, 2012; Essa & Ayad, 2012). However, the previous applications of analytics have disclosed a further research to apply the elaborated analysis and develop more precise prediction models to prevent the drawbacks from the wrong feedbacks to students(Kruse & Pongsajapan, 2012). Therefore, as a preliminary research, this study highlights the need of the deep examinations of current usages and patterns of LMS. Instead of analyzing the individual student level data, the academic course data as a unit of analysis was utilized. We argue that without the thorough analysis on LMS usages and patterns and accurate clustering of the courses, it would not be able to build elaborated prediction models to estimate students' success and failure based on the online behavior records in LMS.

For the rigorousness and thoroughness on data analytics, we employed three methods of clustering analysis: Gaussian Mixture Model, K-Means clustering and Hierarchical clustering. The clustering methods are various and have different strengths. Since using different clustering methods lead different results of clusters, it requires researchers' insights and right interpretations. Therefore, this study attempted to use and present different clustering results from three clustering analysis methods. As mentioned earlier, we utilized LMS dataset to analyze students' virtual learning behaviors and Course Management System (CMS) data to collect the academic course's general information. By using both LMS and CMS data, the clustering analysis of academic courses on the basis of virtual learning environment usage levels and patterns were synergistically performed.

Previous Studies

Academic analytics

The use of Big Data is a mega trend in the current society. Although the Big Data is not a new phenomenon in considering that the large number of data has been incorporated since the invention of World Wide Web in 1989 (Daniel, 2017), dramatical increase of data-accessibility as well as digitalization movements have significantly stimulated to the use of Big Data and data-driven decision-making. The education field is also in the trend of Big Data. Early researchers (Baer & Campbell, 2011; Eynon, 2013; Siemens & Long, 2011) already predicted that the Big Data techniques will make a change of the traditional education system in the educational service, administration, policy, teaching and learning methods. The term "Academic Analytics" reflects such a use of Big Data for the purpose of helping students' academic success. It refers to a tool to respond to increased concerns for accountability in higher education and to develop actionable intelligence to improve student success and learning environment (Campbell et al., 2007; Park, Yu, & Jo, 2016).

Previous studies on the Academic Analytics have reported several innovative cases conducted in higher education. The most frequently cited case was Purdue University's Course Signal (CS). CS is an Early Warning System(EWS) that informs the risk level of each student with green, yellow, or red signal. According to Arnold (2010), academic analytics consider as "a scalable solution to support student success, familiarize students with campus help resources and improve the fail/withdraw rates of large-enrollment, low interaction courses often associated with first-year college attendance". CS works with an algorithm to predict students' success and failure by incorporating not only LMS data but also CMS data such as students' assignments grade, attendance behavior, and past academic performance. As another example, Krumm et al. (2014) has developed an early warning system,

called Student Explorer. This stem aggregates data from an LMS and inform three academic advisors in STEM (Science, Technology, Engineering, and Mathematics) program with frequent updates on students' academic progress and identification of students who need supports. They tracked over 150 individual students across 400 courses and classified students with three types including Encourage(green), Explore(yellow), and Engage(red) based on student percentage points earned, percentage points behind course average, and site visits percentile rank. These systems hold a promising future that helps students' academic performance at the early stage. Krumm et al, (2014) points out the utility of EWS data for "understanding how, when, and why students' academic performance may be declining" (Krumm et al., 2014, p. 117).

As above described, Academic Analytics can be a useful approach to solve the problems (e.g., academic retention and drop-out issues) and to provide more enhanced academic advices such as EWS based on the data that students leave at the Learning Management System as well as Students' Information System or Course Management System. As Siemens and Long (2011) highlighted that the Academic Analytics reveals the role of data analysis at the *institutional level*, in compared to Learning Analytics which require analyzing the interaction among students, instructors, and learning contents. Previous studies that show the examples and approaches of Academic Analytics provide insights how universities can leverage their decision-making by using students' information and their log data in LMS in order to help their learning and academic performances.

Clustering analysis

In the previous section, we argued that Learning Analytics(LA) and Academic Analytics(AA) can contribute to students' learning and academic performance by providing proper feedbacks *immediately*. However, it should be reminded that there are also raising challenges, related to the *usefulness* of such feedbacks and the *accuracy*

of predicting the learning performance based on learners' behavioral data which are mostly relied on their log files left in LMS (Ferguson, 2012). As a result, researchers have attempted to improve the prediction power and accuracy by incorporating diverse approaches. For example, Jo, Park, Kim, and Song (2014) considered the characteristics of academic courses. They compared the prediction model and prediction power in two blended learning courses including 1) an online discussion-based class and 2) a lecture-based class providing regular basis online lecture notes in LMS. Their study suggested algorithms for predicting academic performance should consider different types of classes which incorporate different pedagogical approaches. Another example, Kim, Park, Yoon, and Jo (2016) attempted to develop a prediction model in two blended learning courses that used asynchronous online discussion environment. They could achieve high accuracy of the prediction model and present the possibility of detecting low achievers by incorporating the diverse proxy variables such as TTL(total time spent on LMS), LVF(LMS visit frequency), DVF(Discussion board visit frequency), NOP(Number of postings), LOP(Length of postings), DTV(Discussion time per visit), LIR(LMS visit interval regularity), DIR(Discussion board visit interval regularity), IDC (in-degree centrality), and ODC(out-degree centrality). They observed the accuracy of the prediction model and found increased accuracy trend from the second week (e.g., 70% in course X) to sixth week (88.37%), and to end week (90.7%) in predicting low and high achievers.

In should be noted that the aforementioned cases commonly incorporated the analytics in different blended learning courses and highlighted the importance of selecting the courses to predict the academic performance based on the students' log files in LMS. The prediction models in their studies have limitations to generalize to other blended learning courses which present different usage levels and patterns in LMS. This requires of categorizing the academic courses by their level and patterns of LMS usage and developing the intelligently adopted prediction models which can be more attracted by university decision–makers. Consequently it

is meaningful job to observe the usage levels and patterns of LMS in the institution level (Park & Jo, 2017). Further, it is necessary to categorize the very various blended learning courses into several types which might show different predictors that estimate the students' academic success more powerfully. In this context, Park, Yu, and Jo (2016) selected most representative blended learning courses, 612 out of 4,416 courses, opened during one semester by filtering the inactive courses in LMS with the inclusion and exclusion criteria in using the major LMS functions such as resources, announcement, Q&A, lecture notes, assignment submission, group works, links, discussion forum, quiz, and Wiki. After that, the courses were clustered into four types (Type C: communication or collaboration, D: delivery or discussion, S: sharing or submission, and I: inactive or immature) by using LCA (Latent Class Analysis) technique. While this study reveals major types of blended learning courses in university with their LMS usage levels and patterns, we do consider that the results from classification can be differed by what data-mining techniques and clustering methods are chosen. Also, it is necessary to observe various academic courses in higher education, including extremely active use of LMS or uniquely clustered patterns in regard to incorporating LMS functions in the virtual class, with different view points on different clustering methodologies. Therefore, this study aimed to re-classify the academic courses with three representative clustering methods: Gaussian Mixture Model, K-Means clustering and Hierarchical clustering.

Research Context

The context of this study was a private university located in Seoul, Korea. With the supports of institution for teaching and learning in the university, we collected academic course data of the year of 2013 fall semester. All courses were opened

using Moodle-based virtual learning environment regardless of the course type such as offline and online. Consequently, total 4,416 courses were analyzed at the initial data analysis step. However, since it was revealed that many courses did not use online campus, the exclusion of such non-active courses were performed. Finally, 2,639 courses were observed for this study with 13 variables.

A data set for the analysis was prepared by combining two databases: CMS and LMS. CMS dataset contained course-related information indicating each student's hierarchical categorizations (i.e., graduate VS. undergraduate, mandatory VS. selective, affiliated colleges and department) and LMS dataset included online behavior tracks (i.e., total number of resources, notices, lecture notes, submissions, group works etc.). We integrated CMS and LMS dataset, and these data were divided in general indicators and activity-based indicators. According to Park and Jo (2016), while the general indicators (with number of members, average log-in frequency per person and number of activity items) refer to the values that estimate the *activation* levels of the LMS, (e.g., number of resources, notices, Q&As, lecture notes, task submissions, group works, links, forum postings, quiz, and wikis etc.), activity-based indicators present the pattern of LMS usages. Table 1 shows the total of 13 variables.

In this study, we took these 13 variables including the general and activity-based indicators because the both usage *levels* and *patterns* of LMS should be considered for the clustering analysis. Especially, we chose the average log-in frequency per person (FRE), instead of the simple average log-in frequency because the frequency values are influenced by the number of students in the specific class. ACT (number of activity items) variable was created to see how many activities, relying on the functions in Moodle-based LMS, were utilized in a virtual classroom. Lastly, the rest 10 variables from RES to POS were chosen because these represent each of activity that students and instructors can utilize in LMS.

Table 1. Variable Summary

	No.	Variable name	Variable explanation
General indicator	1	MEM	Number of members
	2	FRE	Average log-in frequency per person
	3	ACT	Number of activity items
Activity-based indicator	4	RES	Number of resources
	5	NOT	Number of notice
	6	QNA	Number of questions and answers
	7	LEC	Number of lecture notes
	8	SUB	Number of task submissions
	9	GRO	Number of group works
	10	LIN	Number of links
	11	POS	Number of discussion forum postings
	12	QUI	Number of quiz
	13	WIK	Number of wikis

Three Clustering Methods

The purpose of this study hold two things. First is to classify the academic courses in a data-driven approach. Second is to compare the clustering analysis methods. Thus, we reviewed three clustering methods: Gaussian Mixture Model, K-Means clustering and Hierarchical clustering. Here we report the process of conducting the clustering analysis as we briefly introduce the characteristics of each model.

Gaussian mixture model

GMM is a probabilistic model that assumes all data are from the mixture of normal distributions. The variables must be numeric since we assume that the data

are from the multivariate normal distribution. The parameters (the proportion of each group, mean vectors, and variance matrix) are estimated by EM algorithm. In general, the number of clusters is very hard to estimate in the clustering analysis. However, we can estimate the optimal number of clusters in GMM using the Bayesian Information Criterion (BIC). We used the R-package "mclust" for GMM. The mclust package in R can estimate not only the number of clusters but also the optimal form of variance matrix. We used the number of clusters from the GMM for the K-means and the hierarchical clustering, too.

K-means clustering

K-means clustering is one of the most popular clustering method because it is very fast to find clusters and very easy to understand. The objective function of K-means clustering is to minimize the sum of within scatters. Basically, it tries to find the k group that minimizes within-cluster sum of squares; therefore, it maximizes the between-cluster sum of squares. Since it uses the squared Euclidean distances among the objects and the cluster centers are defined as the means of objects in each cluster, all variables must be numeric. We used K-means function in R for the analysis.

Hierarchical clustering

Hierarchical clustering method is used for building a hierarchy of clusters from data. Strategies for this clustering fall into two types: agglomerative for "bottom-up" approach and divisive for "top-down" approach. As illustrated in Figure 1, the algorithm finds the nearest two objects and merges them. It repeats this process until all objects are in one cluster. The final results are usually represented by the dendrogram. The hierarchical clustering methods can give different results depending on which distance metric we use between groups.

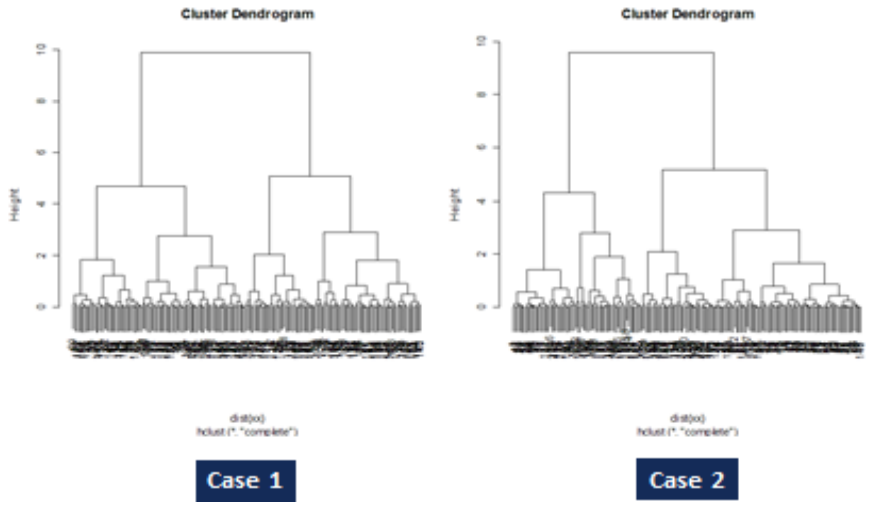

Figure 1. Different hierarchical clustering depending on the distance matrix set-up between groups

There are several distance metrics between groups and we used the "complete-linkage" in our analysis. The "complete-linkage" is the maximum distance between two groups and it is known that the "complete-linkage" can find the compact clusters. We use "hclust" function in R for our analysis.

Results

Descriptive statistics

Before going to the clustering analysis, we checked up on descriptive statistics to find out the distribution of observations. As shown in Table 2, most variables have extremely high values. For example, the maximum values of variables indicate that 596 resources (RES), 176 lecture notes (LEC), 1,612 board postings of group works (GRO), 2,810 discussion postings (POS), and 215 Q&A postings (QNA). These values present extremely high utilization level of few courses.

Table 2. Descriptive statistics of 2,639 courses

Name	Min	Max	Mean	SD	Skewness	Kurtosis
MEM	2	301	33.22	33.66	2.97	13.00
FRE	2	375	39.75	33.01	2.50	11.05
ACT	1	8	2.49	1.30	0.93	0.78
RES	0	**596**	11.87	21.49	12.22	263.56
NOT	0	132	6.64	9.26	3.21	20.15
QNA	0	**280**	2.95	14.25	12.09	183.95
LEC	0	176	3.74	9.69	5.16	51.87
SUB	0	36	0.95	2.82	4.97	32.73
GRO	0	**1612**	17.52	88.42	8.23	91.71
LIN	0	72	0.32	2.57	14.97	312.54
POS	0	**2810**	6.45	75.32	24.44	788.77
QUI	0	**215**	0.61	8.34	17.82	366.00
WIK	0	15	0.01	0.31	42.92	2005.64

On closer inspection, one course which posted 2,810 forum discussion postings was big-sized basic requirement course and students who signed up for class over one hundred. There were 11 groups and they discussed enthusiastically each other, so such very high postings were possible. Next, the other course which had 1,612 group works was the major course of educational technology and the instructor assigned team project during the semester. There were 10 groups and they used group board for team-based learning. Because they uploaded all the related materials for project, opinions and chatting messages in group board, so this high value was also possible. These cases looked as errors but it tells the 'real aspects' of unique courses.

Furthermore, the data were sparse by showing many observations with zero values. The variables from QNA to WIK have zero values for more than 50% of data. We can predict that there will be a single one big cluster with a lot of 'zero' observations. This one big cluster will have all the classes with minimal online

activities. This cluster was not our interest but we were more interested in other clusters of small size and how different they are.

Correlational analysis

Correlations among 13 variables were monitored. As shown in Table 3 and Figure 2, most variables present significant correlations. However, this result was influenced by the high number of cases (n=2,639). A remarkable correlation was found between FRE (average log-in frequency) and ACT (number of activity item) (r=.592, p<.001). ACT variable presented relatively high correlations with other activity-based indicators such as NOT (notice), QNA (question and answer), LEC (lecture note), SUB (task submission), and GRO (group works). This result is no wonder because ACT variable was a general indicator calculated by the sum of activities from NOT, QNA, LEC, GRO, LIN, POS, QUI and WIK.

Table 3. Correlation analysis results (n=2,639)

		1	2	3	4	5	6	7	8	9	10	11	12
1	MEM	-											
2	FRE	.071**	-										
3	ACT	.301**	**.592****	-									
4	RES	.019	.092**	.052*	-								
5	NOT	.394**	**.288****	**.389****	.022	-							
6	QNA	**.419****	.132**	**.257****	.188**	**.233****	-						
7	LEC	.087**	**.452****	**.359****	-.106**	.098**	.046*	-					
8	SUB	-.027	**.475****	**.381****	-.041*	.084**	.039*	**.286****	-				
9	GRO	.123**	**.275****	**.305****	-.001	.197**	.021	.097**	.064*	-			
10	LIN	.098**	**.248****	.243**	-.038*	.083**	.045*	**.260****	.106**	.045*	-		
11	POS	.063*	**.318****	.188**	-.030	.140**	.010	.145**	.062*	.096**	.139**	-	
12	QUI	.016	.186**	.124**	-.020	.033	.031	.064*	.122**	-.015	.159**	.077**	-
13	WIK	.060*	.087**	.113**	.015	.039*	.158*	.053**	.022	-.006	.085**	.075**	.017

*p < .05 **p < .01

An interesting result is that a pair of LEC and SUB shoes relatively higher correlation (r=.233, p<.01) than other pairs among activity-based indicators. Statistically meaningful correlations among activity-based indicators indicate that there might be similar patterns such as the pair of LEC and SUB as we cluster the cases. Therefore, we moved on conducting cluster analysis with different clustering analysis methods.

Figure 2. Paired correlation graph among 13 variables

Gaussian mixture model

First, we conducted a clustering analysis with GMM. As Figure 3 indicates, Mclust found that the best model is three clusters with EEV (ellipsoidal, equal

volume and shape covariance). In the point of three components (clusters), the increase of BIC starts decrease. However, four-cluster model is also close. Thus, we decided to investigate both three-cluster and four-cluster model.

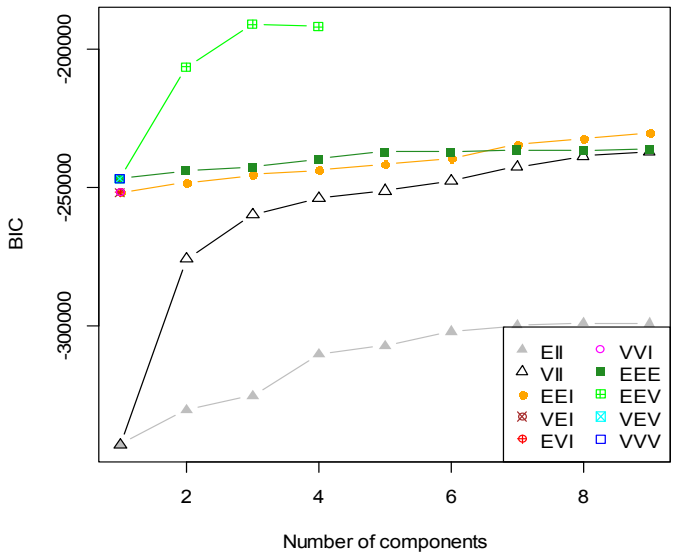

Figure 3. Results of EEV in Mclust

GMM with three clusters

The sizes of three clusters were 212, 2,360, and 67 respectively as seen in Table 4.

Table 4. Clustering table with three clusters

	Cluster 1	Cluster 2	Cluster 3
Number of class	212	2360	67
Mixing probability	0.08068	0.89393	0.02539

We checked the mean vectors (cluster centers) of three clusters. As Figure 4 indicates, cluster 3 (size 67, green line) has the higher mean values (more online activities) and cluster 1 (size 212, black line) is in the middle, and cluster 2 (size 2,360, red line) has the least online activities. On a closer view, cluster 3 has greatly

high value of POS and cluster 1 has high GRO value.

Look inside the clustering table, 2,360 out of total 2,639 classes were included in cluster 2 where having at least online activities. They were inactive classes. Approximately 89% of total class did marginal performance at online campus. On the other hand, cluster 3 was the most active online classes. We can guess that these classes were actively discussed about their topic since both number of forum discussion postings and average log-in frequency per person are quite high. Courses in cluster 1 were also participated in group work much but the average frequency mean was in-between cluster 2 and 3. This cluster was specialized in team project.

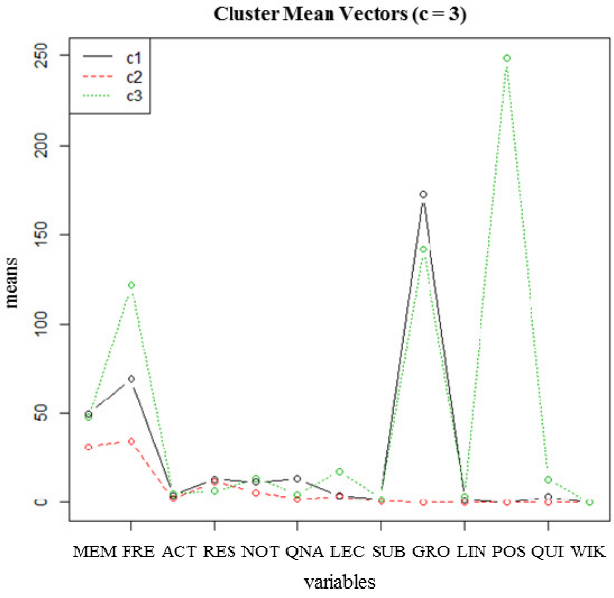

Figure 4. Mean vector plot of three clusters

GMM with four clusters

We divided total classes into four clusters this time. The sizes of four clusters were 71, 2,322, 230, and 16 as sheen in Table 5.

Table 5. Clustering table with four clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of class	71	2322	230	16
Mixing probability	0.02705	0.87962	0.08727	0.00606

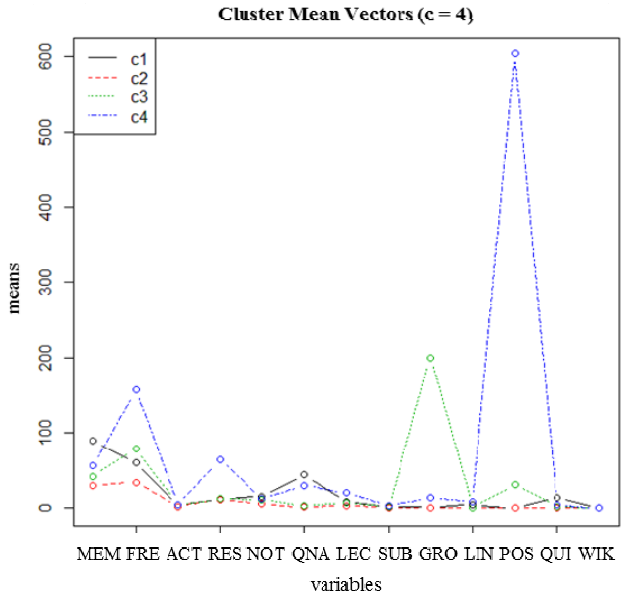

Figure 5. Mean vector plot of four clusters

When reviewing the cluster mean vector plot in Figure 5, cluster 4 (size 16, blue line) has extremely high mean values in POS while cluster 2 (size 2,322, red line) has low mean values in general. Cluster 4 shows the equal appearance with the cluster 3 in GMM with three clustering analysis. Moreover, in common with GMM with three clustering results, 9th variable (GRO) is shown the highest value in cluster 3 (size 230, green line), not the cluster 4 which has higher values in the gross. Courses which involved in cluster 3 were inactive in most of online activities except group works. Newly-drawn cluster 1 (size 71, black line) has the highest MEM value and it represents number of members including an instructor, teaching assistant and students. We are able to call its name, 'big-sized courses'.

The last thing we should observe carefully is that when we clustered total courses into four clusters using GMM, number of courses with highly active in online activities such as forum discussion postings and log-in frequency were decreased from 67 (see Table 4) to 16(see Table 5).

K-means clustering

In addition to GMM, we also performed a clustering analysis using K-Means. As a first step, we analyzed with non-standardized dataset to see overall clusters and compare the results with GMM. However, due to the large scale differences among variables, we also conducted clustering with standardized dataset because we like to see the clustering results when all variables have the similar contributions in distances.

Using non-standardized data

The results of K-means clustering with non-standardized data showed similar results with GMM analysis. But, this process was meaningful because the results identified fewer active online courses.

K-means clustering with three clusters

Most of mean vector values about learners' online behavior were quite similar, similarly low but FRE, GRO, POS variables were distinguished among clusters. Learners who were included in cluster 3 (size 6, green line) in Table 6 logged LMS in the most frequently and wrote up the postings on the forum very much. Cluster 2 (size 71, red line) has high value of GRO which means group works. Cluster 1 (size 2,562, black line) which the most of classes were in has less online action.

Six courses included in cluster 3 are listed on Table 6. They were super active classes in university. As shown in mean vector plot on Figure 6, these courses have high value of log-in frequency (FRE) and forum discussion postings (POS).

Table 6. Clustering table with three clusters

	Cluster 1	Cluster 2	Cluster 3
Number of class	2562	71	6
Mixing probability	0.97082	0.02690	0.00227

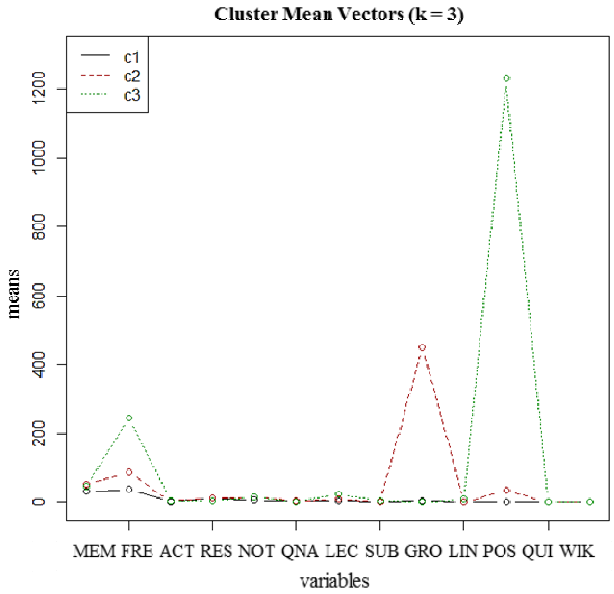

Figure 6. Mean vector plot of three clusters

Table 7. Detailed variable values of cluster 3 courses

No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
255	103	**167**	7	8	71	7	37	3	0	0	**2810**	0	0
894	43	**264**	4	0	0	0	7	14	0	16	**991**	0	0
1299	30	**375**	3	0	0	0	27	0	0	14	**944**	0	0
1403	37	**204**	4	0	0	0	62	1	0	23	**715**	0	0
1630	46	**217**	8	2	22	1	13	12	0	1	**1297**	0	3
2049	18	**245**	4	13	5	1	0	0	0	0	**638**	0	0
M	46	**245**	5	4	16	2	24	5	0	9	**1233**	0	0

K-means clustering with four clusters

When we were partitioning total courses into four clusters, cluster 3 and 4 were somewhat unique. In Figure 7, cluster 3 (size 11, green line) has high value of GRO variable and cluster 4 (size 6, blue line) is shown much online action in FRE and POS variables. As mentioned earlier, students in cluster 4 courses discussed one another constantly and this fact can be proved by FRE and POS. Like the preceding, cluster 3 performed intensive group works. Newly created cluster 2 (size 109) compared to previous results was shown the middle activeness in LMS. In Table 8, six classes included in cluster 4 are exactly the same courses with the cluster 3 in K-means clustering with three clusters analysis (see Table 7).

Table 8. Clustering table with four clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of class	2513	109	11	6
Mixing probability	0.95225	0.04130	0.00417	0.00227

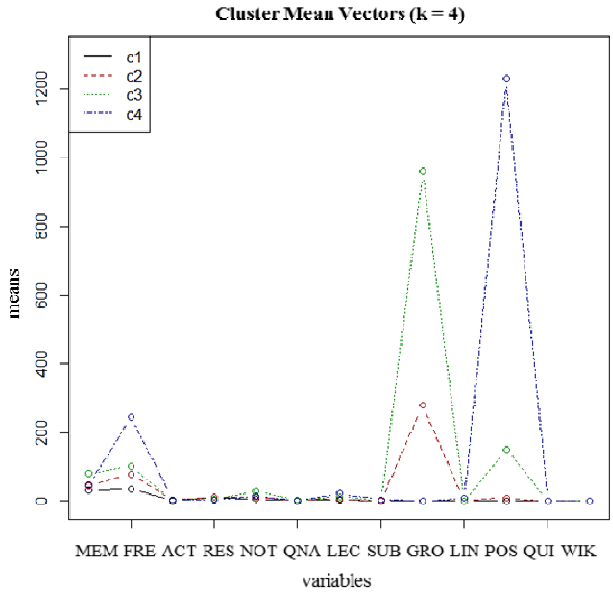

Figure 7. Mean vector plot of four clusters

Using standardized data

Prior to clustering data, we rescaled variables for comparability. So, standardized data was utilized in this step. It showed quite different figures in mean vector plots for the plot of non-standardized dataset. Since K-means uses the squared Euclidean distance, the outliers can affect the clustering results significantly. However, if we use the standardized dataset, then the effect of outliers will be reduced, therefore it is unlikely to see very small sized clusters.

K-means clustering with three clusters

As shown in Figure 8, cluster 2 (size 22, red line) has high mean vector value on the whole. FRE, ACT, LEC, SUB, LIN, POS, QUI and WIK values of cluster 2 were high. Among these, those courses used quiz function very frequently, so QUI was shown excessive activity log in comparison with other clusters.

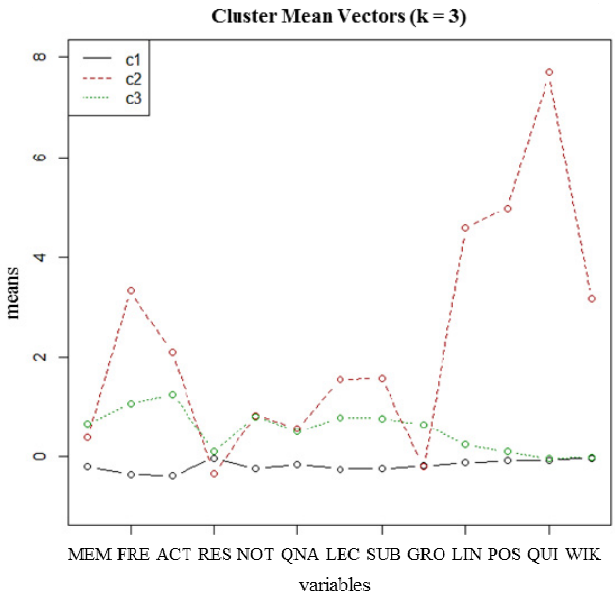

Figure 8. Mean vector plot of three clusters

Table 9. Clustering table with three courses

	Cluster 1	Cluster 2	Cluster 3
Number of class	2030	22	587
Mixing probability	0.76923	0.00834	0.22243

In Table 9, cluster 1 (size 2,030, black line), about 77% of courses contained, was shown the low activeness in general without exception. However, cluster 2 was generally active. Cluster 3 (size 587, green line) was middle-active according to the LMS usage levels, but it had top-of-the-line value in MEM, RES and GRO. In contrast with non-standardized clustering results, these clusters were distinguished by the level of usage, not the unique extreme values.

K-means clustering with four clusters

In Figure 9 and Table10, cluster 3 (size 8, green line) had unusually high mean vector values in QNA, compared to other clusters. Moreover, such MEM, RES and WIK values were also high. We can interpret this situation that there were many members in class, so lots of questions came out together. Another cluster 4 (size 30, blue line) has high action value in FRE, ACT, LEC, SUB, LIN, POS and QUI. In other words, students eagerly participated in LMS in average since the average log-in frequency per person value was the biggest among other cluster. Furthermore, we could assume that the courses provided both a great deal of course-related materials and the grade-related assignment. High values of SUB (number of task submission) and QUI (number of quiz), as well as LEC (number of lecture notes) and LIN (number of URL links) are the evidences. However, cluster 1's (size 1,979, black line) action was minor despite it took most of virtual learning environment courses. Likewise the previous analytic results of cluster 3 (size 587, green line), cluster 2 (size 622, red line) was shown the middle activeness.

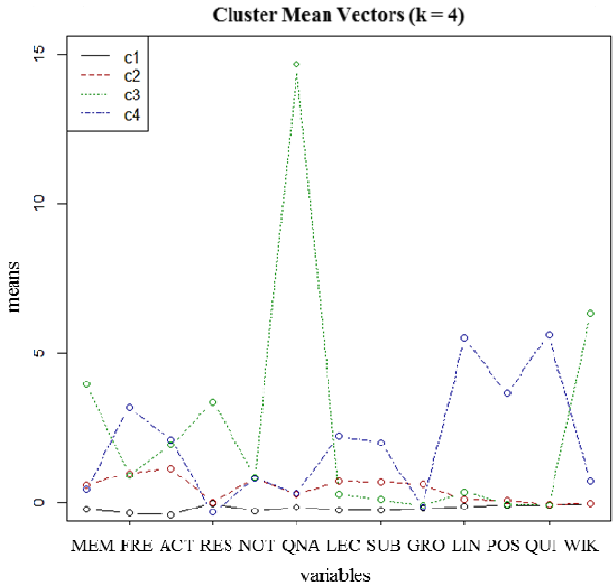

Figure 9. Mean vector plot of four clusters

Table 10. Clustering table with four courses

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of class	1979	622	8	30
Mixing probability	0.74991	0.23570	0.00303	0.01137

Hierarchical clustering

Lastly, we analyzed academic courses with hierarchical clustering method. Standardized dataset was used to clustering.

Hierarchical clustering with three clusters

In Table 11, the result of hierarchical clustering displays an unprecedented appearance. The only 158th class in Table 12 came under cluster 2 (size 1) and a 255th class in Table 13 was included in cluster 3 (size 1). Except those two certain classes, the rest of courses were clustered together in cluster 1 (size 2,637).

Table 11. Clustering table with six clusters

	Cluster 1	Cluster 2	Cluster 3
Number of class	2637	1	1
Mixing probability	0.99924	0.00038	0.00038

Table 12. Detailed variable values of cluster 2 course

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
158	144	93	7	19	8	108	29	0	0	9	30	0	**15**

158th course utilized many activity items (ACT = 7) in moderate way and interestingly used Wiki function in its course. It was the course of economics department. Actually, 15 times was not that huge usage number but as almost the whole courses had not used Wiki (M = .01, SD = .31), this class was chosen for the sole course in cluster 2 because of WIK.

Table 13. Detailed variable values of cluster 3 course

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
255	103	167	7	8	71	7	37	3	0	0	**2810**	0	0

255th class represents extremely high value of forum discussion postings. This course also utilized many activity items (ACT = 7) and specifically in POS, it showed unparalleled usage. It was possible because there were lots of members in class. Every person uploaded 27.28 postings averagely and it would be an acceptable number.

Hierarchical clustering with four clusters

In Table 14, four clusters analytic result was pretty similar with those three clusters hierarchical clustering. Cluster 2 and 3 courses (158th and 255th class) were the same with the previous result. However, newly created cluster 4 (size 3) differed

from the previous one. Three classes out of 2,637 courses had high mean value in RES (number of resources).

Table 14. Clustering table with four clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of class	2634	1	1	3
Mixing probability	0.99811	0.00038	0.00038	0.00114

In Table 15, three courses did not utilize many activities so the variables from SUB to WIK got almost zero value. Specifically, courses had the highest RES values. We can interpret that instructors in these courses chose the resource application instructional method and provided many useful resources for the subject.

Table 15. Detailed variable values of cluster 4 courses

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
514	46	49	2	**596**	0	227	0	0	0	0	0	0	0
1151	84	58	4	**276**	44	19	0	1	0	0	0	0	0
1557	52	83	4	**401**	5	1	29	0	0	0	0	0	0
Mean	60.67	63.33	3.33	**424.33**	16.33	82.33	9.67	0.33	0.00	0.00	0.00	0.00	0.00

Hierarchical clustering with five clusters

In Table 16, Cluster 3 (size 1), 4 (size 1) and 5 (size 3) were same with cluster 2, 3 and 4 in hierarchical clustering with four clusters results. Cluster 2 (size 5) was broken loose from cluster 1 (size 2,629) and five classes in Table 16 were included. These five classes had actively shared useful URL links during the semester as a course material. Mainly, instructors provided references from the web in big-sized courses.

Table 16. Clustering table with five clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Number of class	2629	5	1	1	3
Mixing probability	0.99621	0.00189	0.00038	0.00038	0.00114

Table 17. Detailed variable values of cluster 2 courses

Class No.	MEM	FRE	ACT	RES	NOT	QNA	LEC	SUB	GRO	LIN	POS	QUI	WIK
31	183	49	4	0	10	6	22	0	0	**33**	0	0	0
571	103	59	5	12	6	12	0	2	0	**31**	0	0	0
594	101	52	5	21	24	18	0	1	0	**30**	0	0	0
1243	46	163	7	0	8	8	41	5	0	**48**	201	28	0
1694	55	52	4	0	12	0	33	1	0	**72**	0	0	0
Mean	97.6	75.0	5.0	6.6	12.0	8.8	19.2	1.8	0.0	**42.8**	40.2	5.6	0.0

Discussion and Conclusion

In this study, with the case of academic courses in higher education, we conducted clustering analysis with three different clustering methods. The purpose of this study was to compare these three methods so that researchers can conduct more elaborated and integrated approach when attempting to the interpretation of results. From the literature review, we could organize the general goal, algorithms, characteristics, strengths and weakness of GMM, K-Means and hierarchical clustering analysis (See Table 18). We chose these three methods because each method has different strengths and weakness. As shown in Figure 10 presenting the frequency of words representing these three methods in Google Ngram viewer, while K-means is the most popular method, we can guess that the application of hierarchical clustering method was dominant before 2000 and had been decreased due to its limitations. As described in Table 18, the common purpose of these three methods are to make a group of collection of objects into clusters so that each

clusters are more closely related to one another than the objects assigned to different clusters. However, different algorithms, assumptions, and characteristics of them provide different groups into clusters. Consequently, in our study, each method indicated three to five clusters which present common or contrasting patterns in LMS usage with different group sizes.

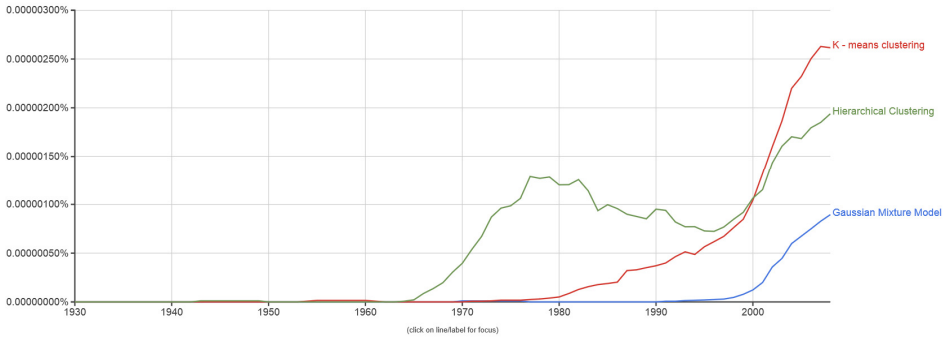

Figure 10. Frequency of words representing three clustering methods.

※ The graph was made with the Google Books Ngram Viewer (Michel et al., 2011), with the standard smoothing of 3.

Figure 11 summarizes the results from this study that clusters academic courses in a higher education. Throughout this comparison on the results, here we discuss several implications, meaningfulness of this study, and suggestions for future study.

First, this study presented considerably unbalanced clusters in academic courses. Three methods *commonly* provided a big cluster representing classes that were almost inactive in terms of LMS usage. On the other hand, each method indicated small portions of active courses that represent different LMS usage patterns. It is meaningful that this result emphasizes the true status quo. The clustering analysis contributes, nevertheless of which method is taken, to filter out the inactive classes in LMS. Throughout this study, we would recommend that educational researchers review the data in two phases. the 1st phase is to filter out the large portion of cluster showing inactiveness. The 2nd phase is to focus on the diverse patterns of active courses by applying several clustering techniques.

Table 18. Comparison of three clustering methods

	GMM	K-means	Hierarchical
Algorithm	Model-based algorithm	Model-free algorithm	
Assumptions	• Observations are samples from mixture distributions	• No probability distribution assumptions exists. • Each algorithm has its own objective function • Many algorithms needs optimization techniques	
Characteristics	• BIC (Bayesian Information Criterion) finds the optimal number of clusters.	• Uses squared Euclidean distance as a dissimilarity measure • Try to minimize the total within scatter	• Bottom-up clustering methods • Starts with *n* clusters and ends with 1 cluster • Find 2 closest objects, merge them to find next closest, merge them until all objects are merged.
Goal	• Group of collection of objects into clusters, such that those within each cluster are *more closely related* to one another than objects assigned to different clusters		
Strengths	• Suggestion of the number of clusters and an appropriate model	• Most well-known clustering algorithm	• Data-driven methods
Weakness	• Various cluster shapes are modeled by different covariance structures.	• The results depend on the initial values • All variables must be quantitative (numeric)	• The results depend on the definition of distance b/w group (e.g., single linkage, complete linkage, group average)

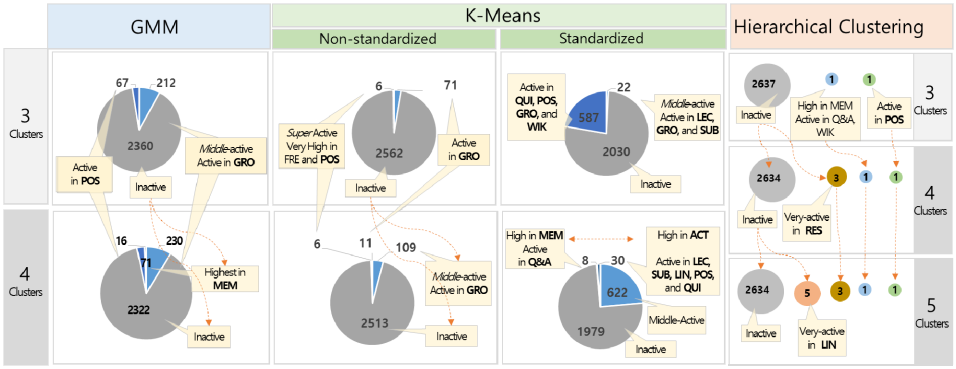

Figure 11. Comparisons on the clustering analysis results from three different methods

Second, this study revealed that certain active courses had *unique* characteristics distinguishing from other courses. As shown in Figure 11, three methodologies revealed several groups of courses that present similar patterns in LMS usage. That is, these methods contributed to this study *in the different manner*. GMM, as an initial step, was essential to check overall clusters of 2,639 academic courses opened during one semester. As classic and the most popular algorithm, K-means with

both non-standardized and standardized dataset contributed to identify prototypical LMS usage patterns by revealing clusters of course utilized forum-based online instruction, quiz-based online instruction, and wiki-based instruction. Hierarchical clustering method was also valuable for the detection of *extreme outlier courses* that revealed resource-based online instruction. Because of hierarchical analytic approach, few outliers could not be included in other cluster naturally but it was left in isolation. This study confirmed that the different strengths of three methodologies leveraged to escalate the effectiveness and robustness of clustering analysis.

Third, this study suggests that educational researchers need to integrate diverse methods to look into the levels and patterns of phenomenon. Clustering analysis is one of useful big-data techniques. Especially, in the education field, clustering can contribute to more personalized and customized educational services. Previous studies attempting clustering analysis, introduced earlier, have incorporated one clustering method. However, as this study shows, three methods result in different clusters that represent different patterns of LMS usages. Even, within the K-Means method, whether or not the variables were standardized led different results. As marked in Figure 11, the standardized model with 4 clusters in K-Means presents a contrast between 8 courses that have high values in MEM (member of course) and 30 courses that have high values in ACT representing diverse uses of LMS function. Both groups utilized LMS actively but very differently. We do believe that such an interesting result was discovered because this study attempted very diverse analytical method.

Finally, although the study has many meaningfulness described above, it has several limitations. For example, we have chosen 13 variables including 3 general indicators, and 10 activity-based indicators which represent the major functions of LMS. However, more variables can be considered to cluster academic courses. While the academic courses in the university consist of blended learning courses, one hundred percent e-learning courses and face-to-face only courses, we did not

consider such a critical classification into the cluster analysis. Therefore, this study suggest that educational policy makers and leaders open the opportunity to explore diverse variables which can impact university's educational services and support for instructors and students.

References

Arnold, K. E. (2010). Signals: Applying Academic Analytics. *Educause Quarterly, 33*(1), n1.

Arnold, K. E., & Pistilli, M. D. (2012). *Course signals at Purdue: Using learning analytics to increase student success.* Paper presented at the Proceedings of the 2nd International Conference on Learning Analytics and Knowledge.

Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning, 4*(2), 17.

Baer, L., & Campbell, J. (2011). Game changers: Education and information technologies. EDUCAUSE. Retrieved March 24, 2014.

Brooks, C., Greer, J., & Gutwin, C. (2014). The data-assisted approach to building intelligent technology-enhanced learning environments *Learning Analytics* (pp. 123-156): Springer.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *Educause Review, 42*(4), 40.

Daniel, B. K. (2017). Overview of Big Data and Analytics in Higher Education *Big Data and Learning Analytics in Higher Education* (pp. 1-4): Springer.

Essa, A., & Ayad, H. (2012). *Student success system: risk analytics and data visualization using ensembles of predictive models.* Paper presented at the Proceedings of the 2nd International Conference on Learning Analytics and Knowledge.

Eynon, R. (2013). The rise of Big Data: what does it mean for education, technology, and media research? : Taylor & Francis.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5-6), 304-317.

Goldstein, P. J., & Katz, R. N. (2005). *Academic analytics: The uses of management information and technology in higher education*: Educause.

Jo, I., Park, Y., Kim, J., & Song, J. (2014). Analysis of online behavior and prediction of learning performance in blended learning environments.

Educational Technology International, 15(2), 137-153.

Kim, D., Park, Y., Yoon, M., & Jo, I.-H. (2016). Toward evidence-based learning analytics: Using proxy variables to improve asynchronous online discussion environments. *The Internet and Higher Education, 30*, 30-43.

Krumm, A. E., Waddington, R. J., Teasley, S. D., & Lonn, S. (2014). A learning management system-based early warning system for academic advising in undergraduate engineering *Learning analytics* (pp. 103-119): Springer.

Kruse, A., & Pongsajapan, R. (2012). Student-centered learning analytics. *CNDLS Thought Papers, 1*(9).

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., . . . Orwant, J. (2011). Quantitative analysis of culture using millions of digitized books. *science, 331*(6014), 176-182.

Park, Y., & Jo, I.-H. (2017). Using log variables in a learning management system to evaluate learning activity using the lens of activity theory. *Assessment & Evaluation in Higher Education, 42*(4), 531-547.

Park, Y., Yu, J. H., & Jo, I.-H. (2016). Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute. *The Internet and Higher Education, 29*, 1-11.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review, 46*(5), 30-32.

Il-Hyun JO

Professor, Dept. of Educational Technology, College of Education, Ewha Womans University.

Interests: Learning Analytics, Social Network Analysis, Knowledge Management, Human Resource Development, Mobile-based Informal Learning in Workplace

E-mail: ijo@ewha.ac.kr

Yeonjeong PARK

Assistant Professor, Dept. of Early Childhood Education (Dedicated Professor, Institue of Teaching an Learning), Honam University

Interests: Mobile and Smart Learning, Socio-cultural Aspects of Learning, Learning and Academic Analytics, Higher Education

E-mail: ypark@honam.ac.kr

Jongwoo SONG

Professor, Department of Statistics, Ewha Womans University

Interests: Regression, Classification, Extreme value theory, Datamining, Computational Statistics

E-mail: josong@ewha.ac.kr

Received: August 30, 2017 / Peer review completed: September 21, 2017 / Accepted: September 22, 2017