Comparison Clustering Performance Based on Moodle Log Mining

C Pradana, S S Kusumawardani, and A E Permanasari

Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia

E-mail: christian.pradana.1@mail.ugm.ac.id, suning@ugm.ac.id, adhistya@ugm.ac.id

Abstract. The use of digital data now saves a lot of information, but still raw and not in the form of knowledge that can be directly seen, therefore we need a data processing to get useful particular understanding from raw data. Data Mining has an important role to process and find useful information from data. This process is also called Knowledge Discovery. Educational Data Mining is a Knowledge Discovery process in the world of education using data mining techniques. One method used in data mining is clustering. By using clustering analysis techniques, data can be grouped into groups without the need for prior knowledge (previous knowledge), so that the data is grouped based on similarity of patterns. This paper compares the K-means, Hierarchical and Louvain clustering methods to see the most appropriate clustering technique in analyzing log activity data in Moodle Learning Management System (LMS). The results of clustering are measured using the Silhouette Coefficient, and then we compare the values and distribution between clusters. In conclusion, Hierarchical clustering produces the highest Silhouette Coefficient value and also this algorithm can detect outlier data as new cluster. Louvain clustering perform very well to find cluster groups in new dataset as the algorithm does not required the number of clusters to be specified before. Louvain clustering can divide more evenly and precision compare to K-means and hierarchical cluster techniques.

1. Introduction

Data is now increasingly important and widely used. Almost all aspects of digital store data in database storage. Data stores a variety of valuable information, but the data still consists of various kinds of diverse information. To find the important information, a data collection technique known as data mining is needed. One area that can use data mining is education. User visits on a page and how long users spend on a page is closely related to user interest in using the website. Web Usage Mining (WUM) is a field that utilizes statistics, machine learning, and data mining algorithms in various types of educational data (1). Various types of variables can be analyzed, for example how long students can complete an assignment and how students interact with e-learning, choosing an analytical method must be done carefully (2). The availability of abundant data raises, so it needs to be able to utilize the information and knowledge. A way to obtain the information from data is using data mining techniques. Therefore, the process of data mining is also called the process of finding knowledge or we know as Knowledge Discovery (KD) (3). The world of education cannot be separated from the use
of technology, from the Learning Management System (LMS), students study report information, to the administration of student data. Modular Object-Oriented Development Learning Environment (Moodle) is a web-based LMS that is commonly used to help learning. All of used of this kind of technology produces very large data, and can be utilized to seek knowledge from these data, therefore Educational Data Mining is needed. The use of data mining techniques on student data also can help the university to do data analysis and determine the promotion strategy plan for marketing to recruit new student (4). By using clustering techniques, the results of grouping students can be obtained, and the results can be used to determine the accurate promotion strategy.

This paper will discuss clustering algorithms for analyzing logs in the Moodle Learning Management System (LMS). Literature review from various aspects from the use of data mining in the field of education, discusses methods, tools, as well as processes commonly used in Web Usage Mining, and also process of finding information through data mining from website data logs in Moodle LMS. The final result of analyzing log data we can search for patterns and explore user groups using LMS.

This paper consists of six parts, in the initial part discussing the role of using and explaining data mining in the field of education. The second part discusses the explanation and description of clustering algorithms and evaluating clustering methods using the Silhouette Coefficient. Henceforth the third section discusses surveys from the literature paper relating to clustering in Educational Data Mining. And in the fourth section discusses methodology and testing of clustering algorithms on the datasets that have been selected in this study. In the fifth section is the result of the research paper, and finally, discussion in section six discusses the conclusion of the results and identification for future work and those related to research development.

2. Background
Data is not always in an ideal condition for analysis, therefore special techniques for data analysis are needed, so it becomes information and knowledge. One of the techniques commonly used is clustering, this technique divides by partitioning a group of data into groups. Data grouped into clusters based on similarity features, and distinct characteristics from other clusters. Before analyzed process, data must first be normalized or original value depend on the case, after the process we can continue analyzed the dataset. There are various types of tools used for Web Usage Mining (WUM), these tools can be used for data normalization, statistical data analysis, after that the results of the data that has been processed can be visualized. In general, there are three processes carried out for WUM, the first is data manipulation, algorithmic analysis, data visualization.

1. Data manipulation, in this section raw data must first be prepared, data must be cleaned (data cleaning), organized (organizing data), and create data formats (creating data). Tools are used to manipulate data are Microsoft excel/google sheets, EDM workbench, Phyton and Jupyter notebook, SQL, logstash.
2. Algorithm Analysis, after data preparation process has been completed, data ready for analyzed process, some tools are used at this stage are Rapid Miner, WEKA, SPSS, KNIME, Orange3, KEEL, Spark MLLib, elasticsearch.
3. Data visualization, after clean data already completely processed, we applied the algorithm and get visual form of data so it is more easily analyzed. Tools used for data visualization include Tableau, D3.js, Orange3 Kibana.

2.1. Clustering Method
Clustering method is a data mining technique that used to analyze data in an unsupervised manner, this technique will divide data into sub data that have been grouped based on similarity relationships. Clustering can make groups of data based on the similarity of data types into one group and separate
them from data that has differences from another group of data. In this paper we focus on three types of techniques, Partition, Hierarchical proposed by Kim and Han(5) and Han et al(6) which have been used frequently in various studies, and Modularity Method proposed by Blondel et al(7) which is used to detect groups in large networks in this study will be compared with Partition and Hierarchical clustering algorithms.

1. Partition: This algorithm will create a group of objects o into a partition c (c <= o) where each partition represents the same object, and which is not the same as the object in another cluster. Each cluster is represented by a centroid.
2. Hierarchical: This hierarchical method will create groups of data, into a cluster tree or hierarchy.
3. Modularity Method is using community detection algorithm, this method assign node as a communities by evaluating how much more densely connected the nodes within the community to compare with other random network, and then maximize modularity score for each community in large network (7).

2.2. Clustering Algorithm

1. K-Means is a clustering algorithm using a partitioning method to analyze data that involves grouping data by partitioning. This algorithm will make the object group o into the c position (c <= o).
2. Hierarchical Clustering is a clustering algorithm that uses an agglomerative hierarchical approach and has good performance on large datasets with multiple large dimensions.
3. Louvain Clustering is clustering algorithm using modular communities for detecting large network, it maximized modularity score and partition size for each community(7).

2.3. Clustering Validation
Cluster validation technique evaluate the performance of a clustering algorithm, consisting of internal and external evaluations. External evaluation is done by comparing the results and ground truth, for example comparing the results of clustering with public class labels to determine how well the cluster technique performs with actual data groups on the available dataset(8). Internal techniques evaluate the similarity of one object in a cluster and measuring the difference between one cluster with another cluster. Silhouette Coefficient (SC) is an internal validity measure that evaluates clustering performance base on the pairwise difference between and within cluster distance(9). The results are between -1 and 1, the closer to 1, the closer to the similarities between entities, conversely, if -1, the less the same between entities.

3. Previous Work
In section three provides an overview various kind of research using data mining methods that are applied to Moodle logs using Web Usage Mining techniques. Research collected from paper journals and conferences. From the literature review it can be seen, certain methods can be used in Moodle logs and the results are used to gain knowledge. Generally, the clustering method is widely used to classify student data based on certain attributes(4)(10)(11)(12)(13). And the results of this grouping can be used for promotion strategies(14) or predictions of student grades. In addition, the Vector Space Model algorithm method can be used to view learning behavior of Learning Management System users(11). Apriori Association algorithm can be used to see the connectedness of data, an example of application in the world of education can be seen in the influence of the relationship between student data and student grades(15). To make predictions of the value of user behavior can be done using the Linear Temporal Logic algorithm, so that the relationship between the two attributes can be seen(16).

Analysis of student access time can be done using the Time Series Cross Section clustering method. By using this method can be seen how many users at each unit of time. So, in the end it can
be seen user behavior based on access time (17)(18). For the initial stage of analysis, the data can be analyzed using statistics (19), to see the characteristics of the data, so that the advanced method can be chosen based on the initial analysis of the data. Linear Regression and Multiple Regression methods can be used to get predictive results of the final grade and the number of students taking a course (20)(21). Taking data values from LMS, then the value is used to determine the classification of student groups based on the scores obtained. Values that do not meet the limit, will get a module that is different from the value reaching the limit. Less value groups will be given a module to help to increase grades. After that, a cluster is created based on the value group. After that the results from the cluster students with the lowest grades are given additional tutoring. From the table below, the literature review grouped by method using the Moodle log dataset.

Table 1. Survey of Clustering Analysis for EDM

| Paper          | Year | Method                                      | Dataset                                                                 |
|----------------|------|---------------------------------------------|-------------------------------------------------------------------------|
| Bovo et al     | 2013 | 10-Fold Cross Validation combine with Expectation maximization hierarchical clustering, Simple K-means, X-means | e-learning training (administrative, on-site training, contact, communication history) and logs from Moodle database |
| Maheswari et al| 2014 | Greedy clustering using belief function     | Collage website log                                                     |
| Pradas et al   | 2015 | Multiple regression Analysis               | Moodle interaction data log in master degree program at Universidad a Distancia de Madrid |
| Gokhan         | 2015 | K-means clustering algorithm               | Student Moodle logs                                                     |
| Mlynarska et al| 2016 | Time series clustering with dynamic time warping | A year of Moodle activity log for 16 computer science courses University College Dublin |
| Alvares et al  | 2016 | Linear Temporal Logic                      | Moodle log data from course at The University of Zaragoza               |
| Dobashi        | 2017 | Time series cross section analysis         | Moodle course log                                                       |
| Estacio et al  | 2017 | Vector Space Model algorithm               | Moodle log data from various courses at Jose Rizal University           |
| Kurniawati     | 2017 | K-means clustering                         | Students data from Institut Sains dan Teknologi Al-Kamal                |
| Hervianti      | 2018 | K-means clustering                         | Students data Universitas Gunadarma                                     |
| Kadoic et al   | 2018 | Statistical analysis included ANOVA, Chi-square test, linear regression | Moodle course log from “Business Decision Making” at University of Zagreb |
| Prastyo et al  | 2018 | K-means clustering                         | Student data from mahasiswa Politeknik Negeri Ujung Pandang             |
| Ikraith-       | 2018 | Association Apriori Algorithm             | Kalam Kudus Student data                                               |
| Informatika    |      |                                             |                                                                         |
4. Methodology
The data taken contained 96,127 logs, generated from 863 users in 139 courses, which were available between August 1, 2017 and December 31, 2017. To find information from these log data, clustering analysis was carried out on several courses. Data consists of columns, index 1, index 2, course, time, user, action, info as shown in Figure 1.

| INDEX 1 | INDEX 2 | COURSE | TIME          | USER  | ACTION           | INFO |
|---------|---------|--------|---------------|-------|------------------|------|
| 0       | -       | C001   | 2017-08-01 19:48:40 | name1 | assignment upload | info1 |
| 2       | 2       | C001   | 2017-08-26 05:27:49 | name2 | resource view    | info2 |
| 3       | 0       | C002   | 2017-09-01 23:36:00 | name3 | assignment view  | info3 |
|         | ...     | ...    | ...           | ...   | ...              | ...  |
| 33587   | 21546   | C003   | 2017-09-30 00:00:28 | name4 | course view      | info4 |
|         | ...     | ...    | ...           | ...   | ...              | ...  |
| 96126   | 19970   | C005   | 2017-12-31 08:43:22 | name5 | resource view    | info5 |

**Figure 1. Moodle Log**

Not all available courses have a view assignment and an upload assignment. The courses are chosen based on the number of activities, interactions between lecturers who give assignments and students who complete the task to see the performance of each clustering algorithm. Four courses with the highest activity were selected in this study, C001 with 137 users and 3873 actions, C002 with 97 users and 5928 actions, C003 with 94 users and 3970 actions, C004 with 79 users and 3067 actions.

**Figure 2. Step for Performing Clustering Evaluation**
First, we extract data from log activity from Moodle LMS as shown in Figure 2. At the pre-processing stage, data collection, feature selection, and data filtering are selected. At the collection stage, data from each user activity on one course is collected to produce a feature vector, which consists of resource view, course view, assignment view, upload assignment. The activity from this dimension was chosen to see every interaction between lecturers in giving assignments and students in completing assignments. Each of these activities becomes the dimension of every user in new data table.

**Table 2. Feature vector generated from data summarization of Moodle log data**

| Dimension         | Description                                                                 |
|-------------------|------------------------------------------------------------------------------|
| assignment upload | Aggregate all uploading file action for each user within the course.         |
| assignment view   | Aggregate all assignment view action for each user within the course.         |
| course view       | Aggregate all course view action for each user within the course.             |
| resource view     | Aggregate all resource view action for each user within the course.           |

**Table 3. Four-dimension data after pre-processing**

| index | user  | assignment upload | assignment view | course view | resource view |
|-------|-------|-------------------|-----------------|-------------|---------------|
| 1     | user1 | 2                 | 9               | 16          | 21            |
| 2     | user2 | 3                 | 13              | 9           | 13            |
| 3     | user3 | 6                 | 10              | 5           | 9             |
| ...   | ...   | ...               | ...             | ...         | ...           |
| 50    | user50| 2                 | 6               | 10          | 23            |
| ...   | ...   | ...               | ...             | ...         | ...           |
| 94    | user94| 6                 | 9               | 52          | 10            |

The study was conducted by using a combination of these four dimensions in Table 2, then the feature selection was made into the dimensions that can be found in Table 3. It became a new dimension that could be included in the calculation of the cluster algorithm for the analysis phase of the test. In the next step, an algorithm comparison test is carried out with and without data normalization to see the comparison result for each algorithm method, to see whether the data normalization has a significant effect on the test results. The function of data normalization is to minimize the value that shows big differences. It will help clustering algorithm to produce more accurate values. The results are documented and displayed in the next chapter in the results section, in the form of tables and graphs.

**5. Result**

The results table in this chapter shows the results of K-means, Hierarchical, Louvain clustering algorithm performance tests on courses C001, C002, C003, C004. Each algorithm is tested with normalization and without normalization on the dataset. The results table is displayed for comparing each other with normalized and non-normalized algorithm.
Table 4. Result of comparing clustering performance

|       | K-Means | K-Means Normalization | Hierarchical | Hierarchical Normalization | Louvain | Louvain Normalization |
|-------|---------|-----------------------|--------------|-----------------------------|---------|-----------------------|
| C001  |         |                       |              |                             |         |                       |
| C1    | 0.480   | 0.497                 | 0            |                             | 0.484   | 0.664                 |
| C2    | 0.673   | 0.687                 | 0            |                             | 0.235   | 0.241                 |
| C3    | 0.397   | 0.498                 | 0.812        |                             | 0.668   | 0.729                 |
| C4    | 0       | 0                     | 0            |                             | 0.365   | 0.590                 |
| C002  |         |                       |              |                             |         |                       |
| C1    | 0.585   | 0.522                 | 0            |                             | 0.545   | 0.339                 |
| C2    | 0.438   | 0.401                 | 0.593        |                             | 0.344   | 0.398                 |
| C3    | 0.465   | 0.375                 | 0.674        |                             | 0.483   | 0.439                 |
| C003  |         |                       |              |                             |         |                       |
| C1    | 0.582   | 0.556                 | 0            |                             | 0.203   | 0.209                 |
| C2    | 0       | 0                     | 0            |                             | 0.547   | 0.400                 |
| C3    | 0.588   | 0.372                 | 0.594        |                             | 0.474   | 0.332                 |
| C004  |         |                       |              |                             |         |                       |
| C1    | 0.502   | 0.475                 | 0.593        |                             | 0.485   | 0.465                 |
| C2    | 0.530   | 0.414                 | 0            |                             | 0.363   | 0.420                 |
| C3    | 0.386   | 0.344                 | 0.674        |                             | 0.550   | 0.463                 |

Table 4 shows on the C001 course, K-Means does not produce a significant difference in results between non-normalized and normalized data. In hierarchical there is a decrease in the value of the silhouette if the data is normalized. The effect of normalization on the Louvain algorithm can be seen in cluster 1 (C1) and cluster 4 (C4), but there is a decrease in value in cluster 3 (C3).

On the C002 and C003 courses as shown in Table 4 there is no significant difference between normalized and non-normalized data. Slightly differences in value of C002 is seen in Louvain clustering where there is a decrease in normalized value in cluster 1 (C1). In course C003, decrease value occurs in K-Means algorithm with normalized in cluster 3 (C3), Louvain with normalized in cluster 2 (C2), cluster 3 (C3). But increase value found in hierarchical with normalized in cluster 4 (C4). The silhouette coefficient course C004, decrease value found in all normalized algorithm, while in Louvain increase value found only in cluster 2 (C2) with normalized data.

Figure 3. Hierarchical clustering scatter plot in course C001
In the Figure 3 shows the results of the hierarchy clustering in course C001 display with scatter plot, detect the outlier data into separate clusters C1 and C2 with less activity in assignment view and upload but high activity in resource view, so that it can be used for detecting lecturers or students who have different or anomaly activities.

**Figure 4.** Comparison K-Means and Louvain clustering in scatter plot in course C001

Louvain algorithm can make the cluster more evenly distributed compared to K-Means algorithm in every course as display in Figure 4. Louvain's algorithm can create new clusters in groups of data that are not detected by the K-means algorithm.
Figure 5. Comparison K-Means and Louvain clustering in scatter plot detail in course C001

Figure 5 above displays, for example user n has ten assignment view features, six resource views, two upload assignments, two course views. By using K-Means algorithm user n is a part of cluster 2 (C2), while using Louvain algorithm user n is a part of cluster 3 (C3). Louvain algorithm detected clusters more evenly, as seen in cluster 3 (C3) and cluster 4 (C4). Louvain algorithm created cluster 3 (C3) with less assignment and resource view activity and cluster 4 (C4) with medium assignment view and less until medium activity resource view, but K-Means algorithm created only two cluster. Normalized and non-normalized data does not become significant to the results of clustering.

6. Conclusion
To find information from a set of data, analysis is needed so that we need analysis technique, data mining can help find information from abundant data set. Cluster analysis can be used to analyze data without the need for previous knowledge. However, choosing the right clustering algorithm determined by knowledge of the data structure, the type of analysis desired, and the size of the dataset being evaluated.

The purpose of this study is to look at the performance of clustering algorithms by comparing the three types of clustering algorithms. The experiment result confirm that Louvain clustering algorithm
performed better than Hierarchical and K-means clustering algorithm. Louvain clustering has an accuracy to detect groups of data, which cannot be detected by K-means clustering. The interesting performance of Hierarchical clustering algorithm, it can point anomaly data or outliers to create cluster. K-means clustering can obtain high silhouette coefficient values on several clusters that have a number of adjacent data.

A future extension of this research will be comparing the performance of clustering on a database with a number of users and a longer usage time.

References
[1] Romero C, Ventura S. Educational data mining: A review of the state of the art. Vol. 40, IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews. 2010. p. 601–18.
[2] Dutt A, Ismail MA, Herawan T. A Systematic Review on Educational Data Mining. Vol. 5, IEEE Access. 2017. p. 15991–6005.
[3] Omar R, Md Tap AO, Abdullah ZS. Web usage mining: A review of recent works. 2014 5th Int Conf Inf Commun Technol Muslim World, ICT4M 2014. 2014;
[4] Hervianti SP. Universitas Gunadarma Menggunakan Metode Clustering. 2010;
[5] Kim M-S, Han J. A particle-and-density based evolutionary clustering method for dynamic networks. Proc VLDB Endow. 2014;2(1):622–33.
[6] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. Data Mining: Concepts and Techniques. 2012.
[7] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10).
[8] Alvarez P, Fabra J, Hernandez S, Ezpeleta J. Alignment of teacher’s plan and students’ use of LMS resources. Analysis of Moodle logs. 2016 15th Int Conf Inf Technol Based High Educ Training, ITHET 2016. 2016;
[9] Dobashi K. Automatic data integration from Moodle course logs to pivot tables for time series cross section analysis. Procedia Comput Sci [Internet]. 2017;12:1835–44. Available from: http://dx.doi.org/10.1016/j.procs.2017.08.222
[10] Estacio RR, Raga Jr RC. Analyzing students online learning behavior in blended courses using Moodle. Asian Assoc Open Univ J [Internet]. 2017;12(1):52–68. Available from: http://www.emeraldinsight.com/doi/10.1108/AAOUJ-01-2017-0016
[11] Kadoic N, Oreski D. Analysis of student behavior and success based on logs in Moodle. 2018 41st Int Conv Inf Commun Technol Electron Microelectron MIPRO 2018 - Proc. 2018;654–9.
[12] Prastyo PH, Saharuna Z, Abduh I. Analisis Data Atribut Mahasiswa untuk Menentukan Strategi Promosi Kampus Analisis Data Atribut Mahasiswa untuk Menentukan Strategi Promosi Kampus Menggunakan Metode Data Mining. 2018;(October).
[13] Sekolah N, Apriori A. Penerapan algoritma apriori untuk menemukan hubungan data murid dengan nilai sekolah. 2018;2(18):7–12.
[14] Kurniuwati I, Indrajit RE, Fauzi M. Peran Bussines Intelligence Dalam Menentukan Strategi Promosi Penerimaan Mahasiswa Baru. 2017;1(2):70–9.
[15] Iglesias-Pradas S, Ruiz-De-Azcárate C, Agudo-Peregrina ÁF. Assessing the suitability of student interactions from Moodle data logs as predictors of cross-curricular competencies. Comput Human Behav [Internet]. 2015;47:81–9. Available from: http://dx.doi.org/10.1016/j.chb.2014.09.065
[16] Lerche T, Kiel E. Predicting student achievement in learning management systems by log data analysis. Comput Human Behav [Internet]. 2018;89(June):367–72. Available from: https://doi.org/10.1016/j.chb.2018.06.015
[17] Riofrío-Luzcando D, Ramirez J, Berrocal-Lobo M. Predicting Student Actions in a Procedural
Training Environment. IEEE Trans Learn Technol [Internet]. 2017;10(4):1–1. Available from: http://ieeexplore.ieee.org/document/7833102/

[18] Appalla P, Kuthadi VM, Marwala T. An efficient educational data mining approach to support e-learning. Wirel Networks. 2017;23(4):1011–24.

[19] Psaromiligkos Y, Orfanidou M, Kytagias C, Zafiri E. Mining log data for the analysis of learners’ behaviour in web-based learning management systems. Oper Res. 2011;11(2):187–200.

[20] Zaffar M, Hashmani MA, Savita KS. Performance analysis of feature selection algorithm for educational data mining. 2017 IEEE Conf Big Data Anal [Internet]. 2017;7–12. Available from: http://ieeexplore.ieee.org/document/8284099/

[21] Kavitha G, Raj L. Educational Data Mining and Learning Analytics Educational Assistance for Teaching and Learning. Int J Comput Organ Trends [Internet]. 2017;41(1):21–5. Available from: http://www.ijcotjournal.org/archive/ijcot-v41p304

[22] Bovo A, Sanchez S, Heguy O, Duthen Y. Clustering moodle data as a tool for profiling students. 2013 2nd Int Conf E-Learning E-Technologies Educ ICEEE 2013. 2013;4(3):121–6.

[23] Maheswari BU, Sumathi P. A new clustering and preprocessing for web log mining. Proc - 2014 World Congr Comput Commun Technol WCCCT 2014. 2014;25–9.

[24] Gökhan A. Profiling {Students’} {Approaches} to {Learning} through {Moodle} {Logs}. Proc {Multidisciplinary} {Academic} {Conference} {Teaching}, {Prague} 2015, {Czech} {Republic} (MAC-{ETeL} 2015) [Internet]. :7. Available from: https://books.google.co.ma/books?id=SkAmCwAAQBAJ

[25] Młynarska E, Greene D, Cunningham P. Time series clustering of Moodle activity data. CEUR Workshop Proc. 2016;1751:104–15.