International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Analysis of Access Log Data on Web Server Using the K-Means Clustering Method

Prasnavira Satria, Arief Wibowo
Study Program of Informatics Engineering, Faculty of Information Technology, Universitas Budi Luhur, Indonesia
Email address: 1611510528 @ student.budiluhur.ac.id, arief.wibowo @ budiluhur.ac.id

Abstract**—** *Web server is an interoperability service that allows data exchange between clients and servers in real-time. Business growth in the current industrial era 4.0 requires companies to operate quickly, precisely, and safely. Companies that have many branch offices in various regions require web server technology to accommodate requests and provide access to company servers. However, not every time, access to a web server reaches a successful status in data traffic. This study aims to group data from the access log on the webserver. The method used is data mining techniques using the k-means algorithm. The results showed the best clustering model was formed at the number of clusters k = 2 with the test value using the Davies Bouldin Index of 0.069. With the results of the study, the company can conduct surveillance on certain types of access, to minimize risk in requests for access to the webserver.*

Keywords**—** *Data Mining, Clustering, K-means, Log File, Web Server*.

I. INTRODUCTION

In business competition, the speed of processing orders from customers is one of the keys to service. During the Covid-19 pandemic, telecommunications companies had more extensive market penetration. However, due to limited operational activities, telecommunications companies must address social restrictions by prioritizing telemarketing services. In business operations, the sales team in charge of marketing and serving retail outlet agents will collaborate with outlets. To support this performance, the sales team uses the company's web application to process orders from retail outlet agents. There are often problems in processing orders such as slow access speeds that cannot access web applications. Those problems can cause the sales team to be suboptimal and the sales team's income statement for branch offices to be delayed.

The sales team needed an analysis of access to log access files on the webserver to find out the problems experienced by the sales team. Log access is a file that records all activities that occur on a web server. Log access files generally have thousands to tens of thousands of data about data access patterns that are successful, unsuccessful, even unnatural. Therefore appropriate analytical techniques are needed. The work of analyzing large amounts of data can be completed by the data mining method. Data mining is an analysis process that occurs in extensive data, intending to know hidden information. Clustering is a data mining method for grouping data that has the same attributes as one class. In the clustering method, one of the algorithms is k-means clustering.

Previous research on log file analysis using irregular expression techniques was applied to provide information about patterns such as the number of page visits [1]. Other studies produce visualization systems to determine the activity of visitors to a website [2]. Other researchers analyzed log files using classification techniques on the Intrusion Detection System (IDS) data [3]. The irregular expression and classification techniques in the study of log file analysis use k-means clustering to help monitor website user access in the form of grouping content [4]. Another study uses k-means clustering with large data processing techniques to produce a model that can detect an attack with a detection probability of up to 99.68% [5]. K-means clustering and association rule techniques are also used to generate behavioral tendencies of visitors to the site. It can help in the optimization process on the website [6].

This research has used k-means clustering to determine the abnormal activity of access that occurs, using data analysis of server log files. The benefit of this research is that it helps to monitor the webserver in reducing access from abnormal activities, as well as failed access.

II. LITERATURE REVIEW

Data Mining is a term used to find hidden information contained in the form of patterns and rules from large data sets so that they are easy to understand. Information obtained from data mining can be used to help decision making and problem-solving [7]. Clustering is a segmentation method in data mining. The clustering method works by grouping data with the same characteristics as one data with other data into one cluster [8].

K-means Clustering algorithm is a non-hierarchical clustering algorithm that can group data into a cluster [9]. The first step in the K-means algorithm is to determine the number of clusters and determine the centroid of each cluster randomly. The second step calculates the distance of each data to the centroid cluster. The third step groups the data closest to the cluster. The fourth step calculates the new centroid value. The second to fourth step is repeated until there is no data transfer between clusters. The calculation of each data's distance to the centroid cluster uses the Euclidean Distance formula. Euclidean Distance calculation uses equation 1.

$$Ca = \sqrt{(x_i - x_{avg})^2 + (n_i - n_{avg})^2} \quad (1)$$

Remaks :
C_a : The distance of data to cluster center
x_i : The value of *x* to i
x_{avg} : The centroid value of the x cluster
n_i : The value of n_i

Prasnavira Satria and Arief Wibowo, "Analysis of Access Log Data on Web Server Using the K-Means Clustering Method," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 3, pp. 56-62, 2020.

n_{avg} : The centroid value of the *n* cluster

Calculation of the new centroid value using equation 2.

$$C_i = \frac{1}{M} \sum_{j=1}^{M} x_j \quad (2)$$

Remarks :
C_i : The centroid of i cluster
M : The amount of data in the cluster
X_j : The *j* data of i cluster

Davies Bouldin Index is a method used to measure the validity of a cluster formed from the clustering algorithm, whether good or not based on the DB Index value that is close to 0 [10]. The first step of the Davies Bouldin Index method is to calculate the distribution of cluster data by centroid using equations 3.

$$Si = \frac{1}{CY_i} \Sigma_{x \in CY_i} \{||x - z_i||\} \quad (3)$$

Remarks:
S_i : The data distribution on centroids
CY_i : The number of members of i cluster
x : The data in clusters
z_i : The centroid of the i cluster

The second step in the DB Index method is to calculate the distance between the centroid cluster. In the second step, the Euclidean Distance formula can be used if the centroid has more than one attribute. If the attribute of a centroid has only one attribute, then the equation 4 is used.

$$M_{ij} = ||z_i - z_j|| \quad (4)$$

Remarks :
M_{ij} : The distance between centroids of i to j
z_i : The centroid of the i cluster
z_j : The centroid of the j cluster

The third step is to calculate the comparison between the distribution of cluster data with the centroid distance between clusters using the equation 5.

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \quad (5)$$

Remarks :
R_{ij} : The comparison of S_i and M_{ij}
S_i : The data distribution from the i cluster
S_j : The data distribution from the j
M_{ij} : The centroid distance between clusters i to j

The fourth step is to calculate the maximum value of R_{ij} using the equation 6.

$$D_i = \max_{i \neq j} Rij \quad (6)$$

Remarks:
D_i : The maximum value of R_{ij}.
R_{ij} : The value distribution of S_i with M_{ij}.

The fifth step is to calculate the DB Index using the equation 7.

$$DB = \frac{1}{k} \sum_{i=1}^{k} D_i \quad (7)$$

Remarks :
DB : The value of DB Index.
D_i : The maximum value of R_{ij}
k : The amount of cluster

III. RESEARCH METHODOLOGY

This research has five stages, including data selection, data preprocessing and data transformation, application of the model with k-means, model testing with the Davies Bouldin Index, and implementation. The purpose of this study is to determine the abnormal activity of the IP address that accesses the webserver. The research phase used can be seen in Figure 1.

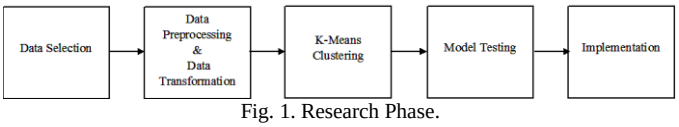


Fig. 1. Research Phase.

The first stage of this research is the selection of data; the data used is the Access Log from the webserver. The second stage of the data will go through preprocessing and transformation to be processed to the third stage. The third stage is the process of calculating k-means clustering, the results of which will be validated in the fourth stage. The fourth step is to validate the results of k-means clustering using the Davies Bouldin index to determine the results. The fifth stage is the implementation of k-means clustering and model testing through applications developed based on PHP language.

IV. RESULTS AND DISCUSSION

A. Data Selection

The data used in this study is the webserver Access Log file for February 2020. The Access log file has a file size of 193KB. The data contained in the Access Log file has several attributes, which can be seen in Table I.

TABLE I. Data Attributes.

No	Attribute	Description
1.	Host	Client's IP address
2.	User-ident	The information about the lient's IP address
3 .	Userid	Client's Username
4.	Time-stamp	Date, Time and GMT Zone
5.	Source path	http method request and source
6.	Status code	The response status of access request
7.	Byte transfered	The byte transfered

In the user-ident attribute containing information from the

Prasnavira Satria and Arief Wibowo, "Analysis of Access Log Data on Web Server Using the K-Means Clustering Method," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 3, pp. 56-62, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

client that accesses, the user-ident will show a dash "-" because, by default, the webserver will make an "off" setting on the identity check feature. The user-id attribute will contain the username of someone accessing it. If the user has not been authenticated or if the requested source path is not protected with a password, it will show a hyphen "-". For the attributes used in this study are time-stamp, status code, and byte transferred.

B. Preprocessing and Data Transformation

Data from the selected attributes will be cleaned in the preprocessing stage so that the data is clean from noise or disturbance. Before preprocessing, the access log file will be opened with a text editor tool with 20,446 rows of known log data. Then the record will be converted into a spreadsheet format without removing the contents of the attributes. The process of changing text to columns from a spreadsheet causes the time-stamp attribute to change into two forms, namely date and time, time zone identification. The results of data transformation can be seen in Figure 2.

host	user-ident	userid	date time	time zone	source path	status code	byte transferred
66.249.71.34	-	-	[01/Feb/2020:00:07:08	+0700]	GET /robots.txt HTTP/1.1	404	385
66.249.72.31	-	-	[01/Feb/2020:00:07:08	+0700]	GET /ads.txt HTTP/1.1	404	382
35.158.114.70	-	-	[01/Feb/2020:00:12:39	+0700]	GET /.env HTTP/1.1	404	379
45.229.53.0	-	-	[01/Feb/2020:00:13:26	+0700]	GET / HTTP/1.1	200	9332

Fig. 2. Results of data transformation.

The next data cleaning process is to remove the symbol "[" in the date-time column so that the data presented becomes valid. In the next stage of data transformation, the fields from date-time will be broken down into two new fields, namely date and time. Date data fields are transformed into values ranging from 1 to 7 according to the order index of days which starts from Monday. The results of data transformation can be seen in Table II.

TABLE II. Results of data transformation of the day.

Day	Value
Monday	1
Tuesday	2
Wednesday	3
Thursday	4
Friday	5
Saturday	6
Sunday	7

For time data transformation is done by taking two digits of the digit unit hours or hours. Thus, the two-digit of the hour unit will be transformed according to the time index from 00 to 23. The final transformation stage is to change the status code to a value ranging from 1 to 5, as shown in Table III.

TABLE III. The transformation result of the status code attribute.

Status Code	Description	Value
404	Not Found	1
400	Bad Request	2
304	Not Modified	3
302	Found	4
200	Sucessful	5

For the byte transferred attribute, no transformation is performed because the data is already in numeric form. In the final results of preprocessing and data, the transformation will produce day index, time index, status code index, and byte transferred, which can be seen in Table IV.

TABLE IV. Results of prerprocessing and data transformation.

No	Date Index	Time Index	Status Code	Byte trans
1.	6	0	1	385 byte
2.	6	0	1	382 byte
3.	6	0	1	379 byte
4.	6	0	5	9,332 byte
5.	6	0	5	9,332 byte
6.	6	0	5	9,332 byte
7.	6	0	5	9,332 byte
8.	6	0	5	9,332 byte
9.	6	1	5	9,332 byte
10.	6	1	5	26 byte
…..	…..	…..	…..	…..
…..	…..	…..	…..	…..
…..	…..	…..	…..	…..
19438.	6	23	2	398 byte

C. K-means Clustering

The clustering modeling stage is completed using the k-means algorithm. The first step is to determine the desired number of clusters. In this study, the determination of the number of clusters k = 2. The two random centroids generated by the system have four attribute values that can be seen in Table V.

TABLE V. Random centroid attribute.

Centroid	Date Index	Time Index	Status Code	Byte trans
0	3	08	1	388
1	4	01	1	381

The random centroid attribute. The next step will be to calculate all the data using the Euclidean Distance equation.

$$Ca(1,0) = \sqrt{(6-3)^2 + (0-8)^2 + (1-1)^2 + (385-388)^2}$$
$$Ca(1,0) = 9.06$$

$$Ca(2,0) = \sqrt{(6-3)^2 + (0-8)^2 + (1-1)^2 + (382-388)^2}$$
$$Ca(2,0) = 10.44$$

$$Ca(3,0) = \sqrt{(6-3)^2 + (0-8)^2 + (1-1)^2 + (379-388)^2}$$
$$Ca(3,0) = 12.41$$

$$Ca(4,0) = \sqrt{(6-3)^2 + (0-8)^2 + (5-1)^2 + (9332-388)^2}$$
$$Ca(4,0) = 8{,}944.00$$

$$Ca(5,0) = \sqrt{(6-3)^2 + (0-8)^2 + (5-1)^2 + (9332-388)^2}$$
$$Ca(5,0) = 8{,}944.00$$

$$Ca(6,0) = \sqrt{(6-3)^2 + (0-8)^2 + (5-1)^2 + (9332-388)^2}$$
$$Ca(6,0) = 8{,}944.00$$

$$Ca(7,0) = \sqrt{(6-3)^2 + (0-8)^2 + (5-1)^2 + (9332-388)^2}$$
$$Ca(7,0) = 8{,}944.00$$

$$Ca(8,0) = \sqrt{(6-3)^2 + (0-8)^2 + (5-1)^2 + (9332-388)^2}$$
$$Ca(8,0) = 8{,}944.00$$
$$Ca(9,0) = \sqrt{(6-3)^2 + (1-8)^2 + (5-1)^2 + (9332-388)^2}$$

Prasnavira Satria and Arief Wibowo, "Analysis of Access Log Data on Web Server Using the K-Means Clustering Method," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 3, pp. 56-62, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Ca(9,0) = 8,944.00

Ca(10,0) = $\sqrt{(6-3)^2+(1-8)^2+(5-1)^2+(385-388)^2}$
Ca(10,0) = 362.10

After the distance between the data with the 0th centroid is known then the data calculation with the 1st centroid is continued with the following example:

Ca(1,1) = $\sqrt{(6-4)^2+(0-1)^2+(1-1)^2+(385-381)^2}$
Ca(1,1) = 4.58

Ca(2,1) = $\sqrt{(6-4)^2+(0-1)^2+(1-1)^2+(382-381)^2}$
Ca(2,1) = 2.45

Ca(3,1) = $\sqrt{(6-4)^2+(0-1)^2+(1-1)^2+(379-381)^2}$
Ca(3,1) = 3.00

Ca(4,1) = $\sqrt{(6-4)^2+(0-1)^2+(5-1)^2+(9332-381)^2}$
Ca(4,1) = 8,951.00

Ca(5,1) = $\sqrt{(6-4)^2+(0-1)^2+(5-1)^2+(9332-381)^2}$
Ca(5,1) = 8,951.00

Ca(6,1) = $\sqrt{(6-4)^2+(0-1)^2+(5-1)^2+(9332-381)^2}$
Ca(6,1) = 8,951.00

Ca(7,1) = $\sqrt{(6-4)^2+(0-1)^2+(5-1)^2+(9332-381)^2}$
Ca(7,1) = 8,951.00

Ca(8,1) = $\sqrt{(6-4)^2+(0-1)^2+(5-1)^2+(9332-381)^2}$
Ca(8,1) = 8,951.00

Ca(9,0) = $\sqrt{(6-4)^2+(1-1)^2+(5-1)^2+(9332-381)^2}$
Ca(9,0) = 8,951.00

Ca(10,1) = $\sqrt{(6-4)^2+(1-1)^2+(5-1)^2+(385-381)^2}$
Ca(10,1) = 355.03

After all, data is calculated the distance from the 0th centroid and the 1st centroid, then the data with the smallest distance results will be grouped into clusters in each centroid. The results of the 0th iteration grouping can be seen in Table VI.

TABLE VI. The results of the 0th iteration grouping.

No.	Centroid 0	Centroid 1	Cluster
1.	9.06	4.58	C1
2.	10.44	2.45	C1
3.	12.41	3.00	C1
4.	8,944.00	8,951.00	C0
5.	8,944.00	8,951.00	C0
6.	8,944.00	8,951.00	C0
7.	8,944.00	8,951.00	C0
8.	8,944.00	8,951.00	C0
9.	8,944.00	8,951.00	C0
10.	362.10	355.03	C1
…..	…..	…..	…..
…..	…..	…..	…..
…..	…..	…..	…..
19438.	18.30	27.89	C1

Then in the 0th iteration it is known that Cluster 0 members are 14,575 and Cluster 1 members are 4,863. After the cluster member is found out in iteration 0, a new centroid is determined. The new centroid calculation uses the following formula:

For 0th Centroid:
Cr0(dateindex) $= \frac{(6+6+6+6+6+\cdots+n)}{14429} = 3.27$
Cr0(timeindex) $= \frac{(0+0+0+0+1+\cdots+n)}{14429} = 11.93$
Cr0(statuscode) $= \frac{(5+5+5+5+5+\cdots+n)}{14429} = 2.03$
Cr0(bytetrans) $= \frac{(9332+9332+9332+9332+9332+\cdots+n)}{14429}$
= 2432.76

For 1st Centroid:
Cr1(dateindex) $= \frac{(6+6+6+\cdots+n)}{5009} = 3.06$
Cr1(timeindex) $= \frac{(0+0+0\ldots+n)}{5009} = 3.12$
Cr1(statuscode) $= \frac{(1+1+1+\cdots+n)}{5009} = 1.40$
Cr1(bytetrans) $= \frac{(382+385+379+\cdots+n)}{5009} = 352.58$

The new centroid value formed can be seen in Table VII.

TABLE VII. New centroid value.

Centroid	Date Index	Time Index	Status Code	Byte trans
0	3.27	11.93	2.03	2,432.76
1	3.06	3.12	1.40	352.58

The next process is the 1st iteration, using a formula Euclidean Distance.

Ca(1,0)= $\sqrt{(6-3.27)^2+(0-11.93)^2+(1-2.03)^2+(385-2432.76)^2}$
Ca(1,0) = 2,047.79

Ca(2,0)= $\sqrt{(6-3.27)^2+(0-11.93)^2+(1-2.03)^2+(382-2432.76)^2}$
Ca(2,0) = 2,050.79

Ca(3,0)= $\sqrt{(6-3.27)^2+(0-11.93)^2+(1-2.03)^2+(379-2432.76)^2}$
Ca(3,0) = 2,053.79

Ca(4,0)= $\sqrt{(6-3.27)^2+(0-11.93)^2+(5-2.03)^2+(9332-2432.76)^2}$
Ca(4,0) = 6,899.25

Ca(5,0)= $\sqrt{(6-3.27)^2+(0-11.93)^2+(5-2.03)^2+(9332-2432.76)^2}$
Ca(5,0) = 6,899.25

Ca(6,0)= $\sqrt{(6-3.27)^2+(0-11.93)^2+(5-2.03)^2+(9332-2432.76)^2}$
Ca(6,0) = 6,899.25

Ca(7,0)= $\sqrt{(6-3.27)^2+(0-11.93)^2+(5-2.03)^2+(9332-2432.76)^2}$

Prasnavira Satria and Arief Wibowo, "Analysis of Access Log Data on Web Server Using the K-Means Clustering Method," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 3, pp. 56-62, 2020.

Ca(7,0) = 6,899.25

Ca(8,0)=
$\sqrt{(6-3.27)^2+(0-11.93)^2+(5-2.03)^2+(9332-2432.76)^2}$
Ca(8,0) = 6,899.25

Ca(9,0)=
$\sqrt{(6-3.27)^2+(1-11.93)^2+(5-2.03)^2+(9332-2432.76)^2}$
Ca(9,0) = 6,899.25

Ca(10,0)=
$\sqrt{(6-3.27)^2+(1-11.93)^2+(5-2.03)^2+(385-2432.76)^2}$
Ca(10,0) = 2,406.79

Calculations with centroid 1 can be seen as follows:

Ca(1,1)=
$\sqrt{(6-3.06)^2+(0-3.12)^2+(1-1.40)^2+(385-352.58)^2}$
Ca(1,1) = 32.70

Ca(2,1)=
$\sqrt{(6-3.06)^2+(0-3.12)^2+(1-1.40)^2+(382-352.58)^2}$
Ca(2,1) = 29.73

Ca(3,1)=
$\sqrt{(6-3.06)^2+(0-3.12)^2+(1-1.40)^2+(379-352.58)^2}$
Ca(3,1) = 26.77

Ca(4,1)=
$\sqrt{(6-3.06)^2+(0-3.12)^2+(5-1.40)^2+(9332-352.58)^2}$
Ca(4,1) = 8,979.42

Ca(5,1)=
$\sqrt{(6-3.06)^2+(0-3.12)^2+(5-1.40)^2+(9332-352.58)^2}$
Ca(5,1) = 8,979.42

Ca(6,1)=
$\sqrt{(6-3.06)^2+(0-3.12)^2+(5-1.40)^2+(9332-352.58)^2}$
Ca(6,1) = 8,979.42

Ca(7,1)=
$\sqrt{(6-3.06)^2+(0-3.12)^2+(5-1.40)^2+(9332-352.58)^2}$
Ca(7,1) = 8,979.42

Ca(8,1)=
$\sqrt{(6-3.06)^2+(0-3.12)^2+(5-1.40)^2+(9332-352.58)^2}$
Ca(8,1) = 8,979.42

Ca(9,0)=
$\sqrt{(6-3.06)^2+(1-3.12)^2+(5-1.40)^2+(9332-352.58)^2}$
Ca(9,0) = 8,979.42

Ca(10,1)=
$\sqrt{(6-3.06)^2+(1-3.12)^2+(5-1.40)^2+(385-352.58)^2}$
Ca(10,1) = 326.62

Based on the results of calculations with the new centroid, then the data with the smallest distance value will be grouped into clusters of each centroid. The results of the 1st iteration grouping can be seen in Table VIII.

TABLE VIII. The results of the 1st iteration grouping.

No	Centroid 0	Centroid 1	Cluster
1.	2,047.79	32.70	C1
2.	2,050.79	29.73	C1
3.	2,053.79	26.77	C1
4.	6,899.25	8,979.42	C0
5.	6,899.25	8,979.42	C0
6.	6,899.25	8,979.42	C0
7.	6,899.25	8,979.42	C0
8.	6,899.25	8,979.42	C0
9.	6,899.25	8,979.42	C0
10.	2,406.79	326.62	C1
…..	…..	…..	…..
…..	…..	…..	…..
…..	…..	…..	…..
19438.	2,034.79	49.67	C1

After the data is grouped into clusters, the next step is to calculate the new centroid value with the following results:

0th Centroid:
Cr0(dateindex) $= \frac{(6+6+6+6+6+\cdots+n)}{3304} = 3.93$
Cr0(timeindex) $= \frac{(0+0+0+0+1+\cdots+n)}{3304} = 11.75$
Cr0(statuscode) $= \frac{(5+5+5+5+5+\cdots+n)}{3304} = 4.99$
Cr0(bytetrans) $= \frac{(9332+9332+9332+9332+9332+9332+\cdots+n)}{3304}$
$= 9280.88$

1st Centroid:
Cr1(dateindex) $= \frac{(6+6+6+6+\cdots+n)}{16134} = 3.07$
Cr1(timeindex) $= \frac{(0+0+0+1+\cdots+n)}{16134} = 9.23$
Cr1(statuscode) $= \frac{(1+1+1+5+\cdots+n)}{16134} = 1.23$
Cr1(bytetrans) $= \frac{(382+385+379+26+\cdots+n)}{16134} = 384.55$

If the number of cluster members does not change, then the k-means calculation process has been completed and produces the desired clustering results in the form of how many members are per cluster. In this study, iteration will take place up to 4 times. The results of the modeling formed are cluster 0 containing 3079 members, and cluster 1 containing 16359 members. Cluster specifications can be seen in Table IX.

TABLE IX. Cluster spesification.

Cluster	Status Code	Description	Frequency	Byte Trans Average
0	200	Successful	3079	9,751.39 byte
1	200	Successful	942	846.58 byte
	302	Found	52	1,631.98 byte
	400	Bad Request	577	398.51 byte
	404	Not Found	14788	378.58 byte

D. Model Testing

After the calculation process is complete, the final centroid will be obtained to calculate the Davies Bouldin Index value. The first step is to calculate the value of the data distribution of the cluster against the final centroid using the last iteration data and the last centroid with the following formula:

Prasnavira Satria and Arief Wibowo, "Analysis of Access Log Data on Web Server Using the K-Means Clustering Method," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 3, pp. 56-62, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

0^{th} cluster data distribution:

$S_0 = \frac{(419.56+419.56+419.56+419.56+419.56+419.53+\cdots+n)}{3079}$

$S_0 = 1225.43$

1^{st} cluster data distribution:

$S_1 = \frac{(37.64+34.75+40.54+392.47+\cdots+n)}{16359}$

$S_1 = 79.16$

The second step is to calculate the distance between the centroid cluster with the Euclidean Distance formula.

$M_{01} = \sqrt{(3.96-3.08)^2+(11.56-9.30)^2+(5-1.28)^2+(9751.39-418.35)^2}$

$M_{01} = 9333.05$

The third step is to calculate the comparison between the distribution of data with the distance between centroids with the following formula:

$R_{01} = \frac{(S0+S1)}{M01}$

$R_{01} = \frac{(1225.43+79.16)}{9333.05}$

$R_{01} = 0.13978$

The fourth step is to calculate the maximum value of each R_{ij} that exists with the following formula:

$Di = \max(R_{01})$

$Di = 0.13978$

The final step is to calculate the value of the DB Index with the following formula:

$DB = \frac{1}{2}\ (0.13978)$

$DB = 0.069$

From the calculation of the Davies Bouldin Index value on the results of clustering with the number of clusters (k) = 2, it produces a DB Index value of 0.069. As a comparison of performance, cluster forming experiments were carried out starting from k = 3 to k = 5. DB Index values from the experimental results can be seen in Table X.

TABLE X. The Davies Bouldin Index comparison.

k	DBI
2	0.069
3	0.137
4	0.636
5	2.021

Based on the comparison of DB Index from Table 9, it is known that the best modeling of clusters is at the number of k = 2 with DB Index value reached 0.069.

E. Implementation

After the modeling and testing phase is completed, the final work of this research is to build the system in the implementation phase. Based on the results of the data transformation in Figure 2, the logs that have been stored in a spreadsheet format will be imported into the system. The import results can be seen in Figure 3.

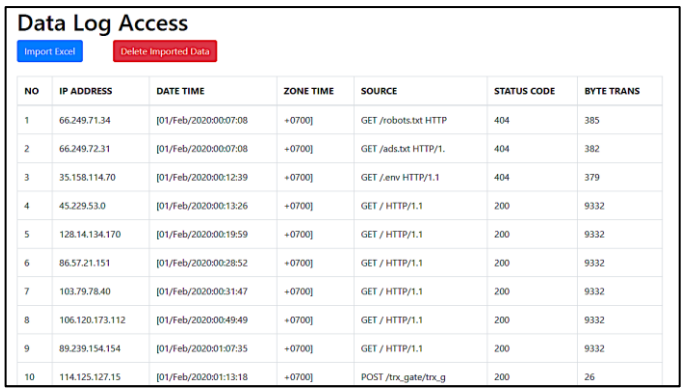


Fig. 3. The results of data transformation in the application

Then the data that has been imported will be preprocessed. The preprocessing process and final results of the preprocessing process that has been processed in the application built can be seen in Figure 4 and Figure 5.

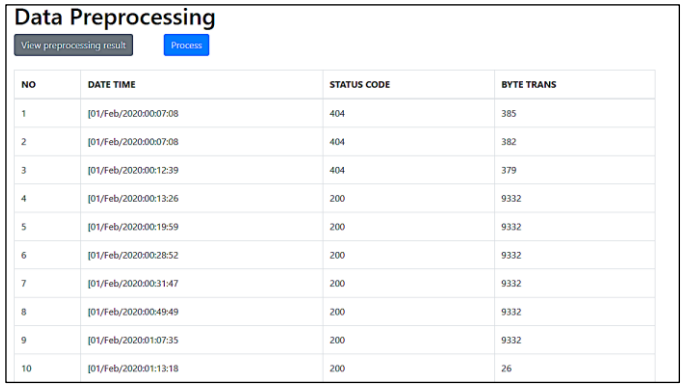


Fig. 4. Data pre-processing in the application.

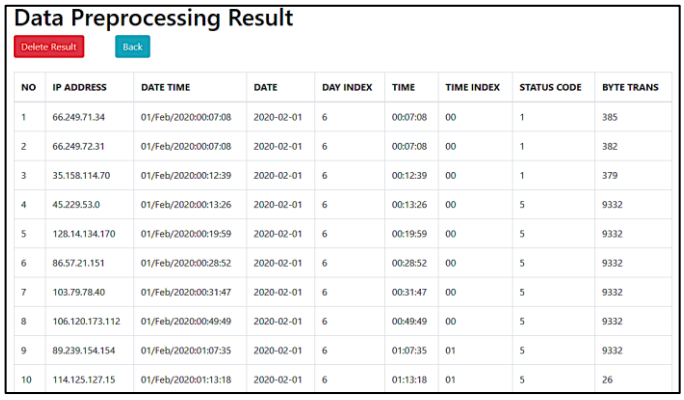
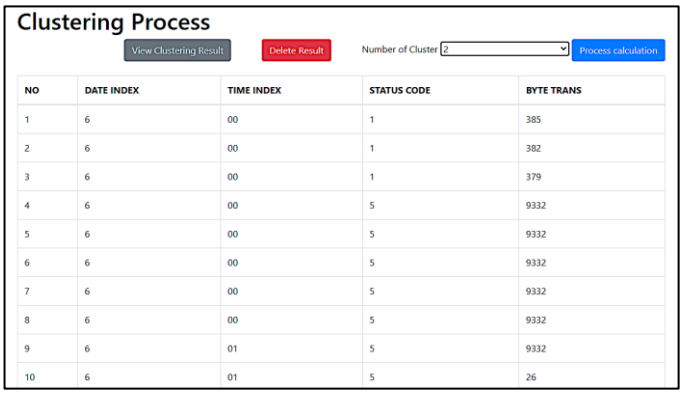


Fig. 5. Results of data pre-processing in the application.

Prasnavira Satria and Arief Wibowo, "Analysis of Access Log Data on Web Server Using the K-Means Clustering Method," *International Research Journal of Advanced Engineering and Science*, Volume 5, Issue 3, pp. 56-62, 2020.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

After preprocessing is completed, the next stage is the clustering with k-means. The clustering with k-means can be seen in Figure 6.



NO	DATE INDEX	TIME INDEX	STATUS CODE	BYTE TRANS
1	6	00	1	385
2	6	00	1	382
3	6	00	1	379
4	6	00	5	9332
5	6	00	5	9332
6	6	00	5	9332
7	6	00	5	9332
8	6	00	5	9332
9	6	01	5	9332
10	6	01	5	26

Fig. 6. Results of data preprocessing in the application.

Afer calculation completed, then determination of the centroid value is done randomly and can be seen in Figure 7.

RANDOM CENTROID

#	DAY INDEX	TIME INDEX	STATUS CODE	BYTE TRANS
0	3	08	1	388
1	4	01	1	381

Fig. 7. Random Centroid.

In Figure 8 can be seen the value of Davies Bouldin Index and cluster specifications formed.

RANDOM CENTROID

#	DAY INDEX	TIME INDEX	STATUS CODE	BYTE TRANS
0	3	08	1	388
1	4	01	1	381

LAST CENTROID

#	DAY INDEX	TIME INDEX	STATUS CODE	BYTE TRANS
0	3	11	5	9754
1	3	9	1	418

NUMBER OF CLUSTER MEMBERS

ITERATION	Cluster 0	Cluster 1
0	14575	4863
1	3305	16133
2	3101	16337
3	3079	16359
4	3079	16359

Davies Bouldin Index score **0.069944301628106**

The conclusions from the results of k-means clustering are:

Export Excel

Cluster 0 : with statuscode 200 appears 3079 times, with an average of bytetrans 9751.39 byte
Cluster 1 : with statuscode 200 appears 942 times, with an average of bytetrans 846.58 byte
Cluster 1 : with statuscode 302 appears 52 times, with an average of bytetrans 1631.98 byte
Cluster 1 : with statuscode 400 appears 577 times, with an average of bytetrans 398.51 byte
Cluster 1 : with statuscode 404 appears 14788 times, with an average of bytetrans 387.58 byte

Fig. 8. The results of k-means clustering modeling.

From the results of the modeling that can be done is to supervise the members of cluster 1 because it has a status code 404 (Not Found), the cause of this status can be traced by advanced analysis methods. Supervision also needs to be done on cluster 0, which has status code 200 (Successful), because the average size of data sent is 9,751 bytes. These facts being an indication of unusual data exchange because the average size of the sent data is 2 KB.

V. CONCLUSION

Based on the study results, it can be concluded that k-means clustering can be used to find out abnormal activity in the access log of the webserver using developed application. The number of clusters k = 2 and the amount of data as much as 19,438 obtained the value of Davies Bouldin Index of 0.069, which is the best clustering model to determine the type of abnormal activity from the data log access web server.

REFERENCES

[1] Yogi, I. Ruslianto, and S. Bahri, "Analisa Log Web Server Untuk Mengetahui Pola Perilaku Pengunjung Website Menggunakan Teknik Regular Expressions," *J. Komput. dan Apl*, vol. 07, no. 01, pp. 120–130, 2019. *(Published in Bahasa)*

[2] R. Andriani, E. S. Pramukantoro, and M. Data, "Pengembangan Sistem Visualisasi Access Log untuk Mengetahui Informasi Aktivitas Pengunjung pada Sebuah Website," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 6, pp. 2104–2112, 2018. *(Published in Bahasa)*

[3] D. Mongkareng, N. A. Setiawan, and A. E. Permanasari, "Implementasi Data Mining dengan Seleksi Fitur untuk Klasifikasi Serangan pada Intrusion Detection System (IDS)," *Citee*, no. gambar 2, pp. 314–321, 2017. *(Published in Bahasa)*

[4] T. A. Al-Asadi and A. J. Obaid, "Discovering similar user navigation behavior in web log data," *Int. J. Appl. Eng. Res*, vol. 11, no. 16, pp. 8797–8805, 2016. *(Published in Bahasa)*

[5] F. Ridho and A. A. Kusuma, "Deteksi Intrusi Jaringan dengan K-Means Clustering pada Akses Log dengan Teknik Pengolahan Big Data," *J. Apl. Stat. Komputasi Stat*, vol. 10, no. 1, p. 53, 2019. *(Published in Bahasa)*

[6] A. Fauzanu, E. Darwiyanto, G. Agung, A. Wisudiawan, F. Informatika, and U. Telkom, "Analisis Web Usage Mining Menggunakan Teknik K-Means Clustering Dan Association Rule (Studi Kasus : www.owlexa.com)," *e-Proceeding Eng*, vol. 4, no. 2, pp. 3284–3291, 2017. *(Published in Bahasa)*

[7] M. N. V. Waworuntu and M. Faisal Amin, "Penerapan Metode K-Means Untuk Pemetaan Calon Penerima Jamkesda," *Klik - Kumpul. J. Ilmu Komput*, vol. 5, no. 2, p. 190, 2018. *(Published in Bahasa)*

[8] R. Maulana, "Web Usage Mining Menggunakan K-means untuk Mengetahui Kecenderungan Akses Pengguna (Studi Kasus: ganto.co), " Jurnal Vokasional Teknik Elektronika dan Informatika (VOTEKNIKA), vol. 6, no. 2, 2018. *(Published in Bahasa)*

[9] D. S. Hermawan, Syaifuddin, and Diah Risqiwati, "Analisa Real-Time Data log honeypot menggunakan Algoritma K-Means pada serangan Distributed Denial of Service," vol. 2, no. 5, pp. 541–552, 2018. *(Published in Bahasa)*

[10] S. Rustam, "Analisa Clustering Phising Dengan K-Means Dalam Meningkatkan Keamanan Komputer," *Ilk. J. Ilm*, vol. 10, no. 2, pp. 175–181, 2018. *(Published in Bahasa)*