(IJID) International Journal on Informatics for Development,
Vol. 6, No. 1, 2017. Pp 17-21

The Mapping of Access Point Workloads at UIN Sunan Kalijaga Based on Log Analysis using K-Means Algorithm

Razendra Bintang Kharisma
Department of Informatics
Universitas Islam Negeri Sunan Kalijaga
Yogyakarta, Indonesia
bintangrazen@gmail.com

Ahmad Subhan Yazid
Department of Informatics
Universitas Islam Negeri Sunan Kalijaga
Yogyakarta, Indonesia
Yazid.anfalah@gmail.com

***Abstract*— This study aims to map the use of access points based on the log of usage in several locations in the UIN Sunan Kalijaga. The analyzed data are access point records taken three times a day for one week working hours. The data obtained was processed and clustered using the K-Means algorithm. There are five clusters obtained from the processed data, each of which divides the number of access loads from an access point. The results of clustering provide a recommendation on which locations need to be added to access points that it can improve institutional user services.**

Keywords: Access loads, Cluster, K-means, Log, Number of users.

I. INTRODUCTION

Today's need for internet is very high, especially in universities. The entire academic community uses the internet to meet each of their needs, such as: finding lectures, accessing e-learning, or building networks. Therefore, providing access points as a connecting device to the internet by universities is important.

To support effective and efficient service provision, a university needs to analyze web server log access data. This analysis can further improve the effectiveness of a site, provide better communication services, and can also increase certain targets for a user group [1]. Log data analysis requires a method that can describe and predict the behavior of internet users. The method used in this study is data clustering which refers to user log data stored on the server web log.

This study uses the K-means algorithm in clustering. This algorithm is the most commonly used because it has the simplicity stages compared to other algorithms. However, this algorithm is also very reliable for handling data with numerical attributes [2].

II. ACCESS POINT AND LOG FILE

Access Points are hardware devices that allow other wireless devices (such as laptops, cellphones) to connect to a network using Wi-Fi, Bluetooth or other standard devices. Wireless Access points are generally connected to routers via a wired network and can be used to send data between wireless devices and cable devices on a network. Access-Point functions to convert radio frequency signals into digital signals that will be channeled through cables, or channeled to other WLAN devices by being converted back into radio frequency signals [3].

Figure 1. CISCO's Access point

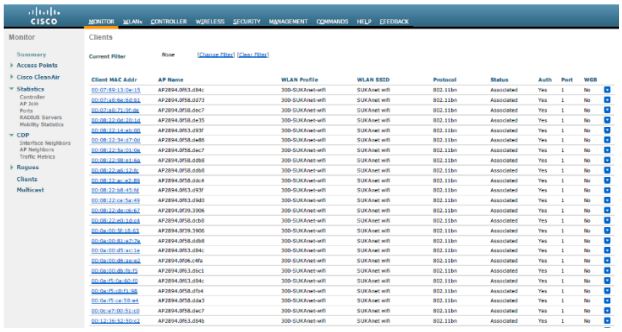
The Access Point used at Sunan Kalijaga State Islamic University is CISCO. It is a global company in the field of telecommunications based in San José, California, United States. The CISCO access point can accommodate a maximum of 128 clients depending on the data communication.

Log is a record of all activities an application is running. Log-File is a file that lists actions, events (activities) that have occurred in a computer system. For example, the web server has a list of log files for each request (request) from the browser that is addressed to the server. With the Log file analyzer, it is possible to know things like where the visitor came from, how often they returned, and how they navigated to the website [4].

A data that can be analyzed from the log of a network device file is Media Access Control Address (Mac Address). MAC Adsress is a network address that is used as a unique identification that is owned by each computer network card, or switch, or router, or access point, or anything that may be connected to the network. With this capability, anyone who accesses a computer network can be identified.

III. K-MEANS ALGORITHM

K-means is an algorithm in data mining that can be used to cluster data. Clustering is the distribution of data into several



groups / clusters that have a similarity. Clustering is part of a data mining study that involves several stages that are carried out before the data is interpreted, namely: data retrieval, preparation, and preprocessing [5].

Clustering of data with the K-means algorithm is carried out with the following stages [6]:

- Determining the number of groups / clusters
- Calculating the group center (centroid / average) of the data in each group. If *M* expresses the amount of data in a group, *i* denotes the *i*-feature in a group, and *p* denotes the dimension of data, then the equation for calculating the feature-centroid i uses the following equation:

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

  equation (1) is done as much as *p* dimensions from *i*=1 to *i*=*p*.

- Calculating the distance of data to each cluster. A method that can be used to measure the distance of data to the center of the group is by Euclidean formula with the following equation:

$$C_a\sqrt{(x_i - x_{avg})^2 + \cdots + (n_i - n_{avg})^2} \qquad (2)$$

With x_i is the first data and x_{avg} is the center point of the first cluster while n_i is the n[th] data and n_{avg}navg is the center point of n[th] data.

- Allocating each data to the nearest centroid / average.

IV. IMPLEMENTATION

The implementation of this study was carried out with the following scheme:

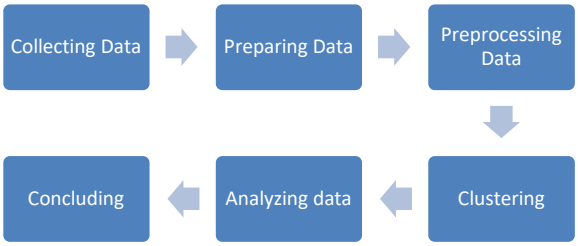


Figure 2. Research implementation scheme

A. Collecting Data

The study was conducted at UIN Sunan Kalijaga by taking log data on internet access from several access point locations. Data is taken three times in one day for one week working hours (5 days), namely morning (09.00-10.00), afternoon (12.00-13.00), and afternoon (15.00-16.00). The raw data obtained is in the format * .pdf and * .xls. In each table there are nine columns. Each column contains the Client MAC Address, AP Name, WLAN Profile, WLAN SSID, Protocol, Status, Auth, Port, WGB attributes. The amount of data collected is 30895 rows, with each row having nine attribute columns, as shown below:

Figure 3. Result of data collection

B. Preparing Data

The data that has been collected is grouped into five folders that are marked as days and each of them has three folders which share data based on access times, namely morning, afternoon and evening. The data is then imported into the MySQL database based on the day and time. For example, kamis_pagi table contains internet log data on Thursday's morning, kamis_siang table contains internet log data on Thursday's afternoon, kamis_sore table contains internet log data on Thursday's evening, and so on.

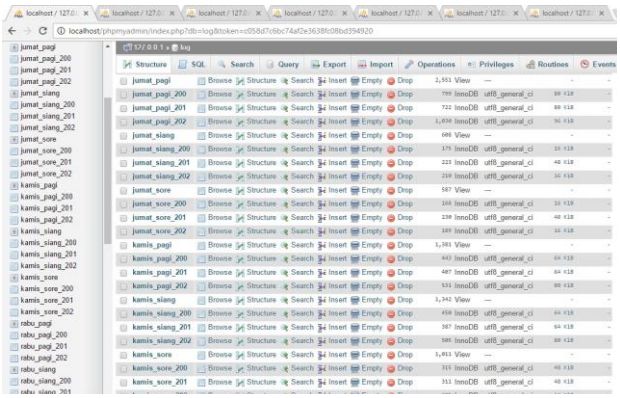


Figure 4. Data in MySQL database

C. Preprocessing data

Data preprocessing is a stage that is carried out to convert or process raw data into data that is ready to be analyzed. This stage aims to eliminate data containing empty data (NULL) and noise (N/A). Data cleaning is very important because the accuracy is the basis for the resulting the consistent data, correct in its format, no duplication, and in accordance with what is needed. There is a difference between raw data and preprocessing results data shown in the following table:

TABLE I. DIFFERENCES IN DATA BEFORE AND AFTER THE CLEANING PROCESS

Before	*Missing Value*	*Noise*	After
30895 *records*	45 *records*	5 *records*	30845 *records*

The attributes needed in the clustering stage are the client mac address and AP name. Therefore seven other attributes are deleted from the database. Furthermore, the original data is transformed into the initial data so it can be identified in the clustering stage. This step adds the attributes of the day, time, location, and changes the initial scenario "A user accesses through an access point". become "One access point accessed by several users".

The following is a comparison of the number of data records in the raw data and the data ready to process:

TABLE II. COMPARISON OF RAW DATA READY FOR PROCESSING

Amount of Raw Data	Amount of Raw Data Field	Amount of Ready Data	Amount of Ready Data
30895 *records*	2 *fields*	1835 *records*	*4 fields*

D. Clustering data: K-Means

The stages carried out in clustering using the K-Means algorithm are as follows:

1) Determine the number of clusters. In this study, 5 clusters were determined based on the characteristics of existing data.

2) Determine the center point of the cluster or centroid. Centroid is the difference value in calculating the distance between data to each cluster. Usually, centroid determination is done randomly. In this process, the researcher creates an initial variable that determines the access load of an access point. The value of the access load starts from the range of zero values (0) to one hundred (100) with very low, low, normal, high, and very high variables.

3) Calculate the distance of data to each cluster with the Euclidean formula. Calculation is done with Ms.'s help Excel as in the following picture:

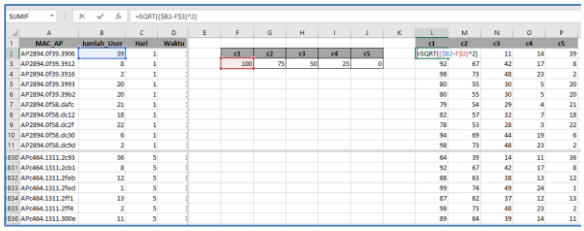


Figure 5. Euclidean formula calculation

4) Allocate data into clusters. Allocation of data is based on the results of the distance between data to each cluster. If the distance value between all clusters is the smallest, it will be given a true value and the other four clusters will be assigned a false value.

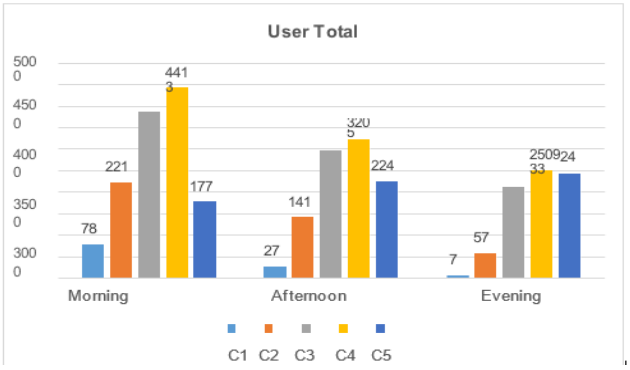
5) Repeat steps 3 and 4 If there is still data that moves groups or if there is a change in the centroid value above the specified threshold value, or if the change in the value of the objective function used is still above the specified threshold value. In this clustering process, researchers repeat fifteen (15) times. In the 15th iteration, the value of the 15th cluster data center point is the same as the cluster center point value in the previous iteration (14th iteration), then the iteration does not need to be continued

V. RESULT AND ANALYSIS

The results of clustering (table) show that in general the distribution of access point points at UIN Sunan Kalijaga is evenly distributed, even though there are several access points that exceed 128 users.

TABLE III. MAXIMUM AND MINIMUM USER ACCESS PER CLUSTER

Cluster	*Centroid Value*	Amount of User Access of each AP		AP Total Points (5 days)	User Total (5 days)
		Max	Min		
C1	86,84615	144	74	13	1129
C2	56,02667	69	47	75	4202
C3	36,17409	46	29	247	8935
C4	20,33534	28	14	498	10127
C5	6,439122	13	1	1002	6452



Meanwhile, if observed from the access point access time, information is obtained that the most internet usage is in the morning, as shown below:

Figure 6. Graph of the number of users based on access time

Then, if analyzed based on the location of the access point, the spread of access points with the most workload (cluster 1) can be seen in the following table:

TABLE IV. LOCATIONS INCLUDED IN THE FIRST CLUSTER

Location	Day	Time	Amount of AP	User Total
Adab Faculty	Wednesday	Morning	1	75
Sceince and Technology Faculty	Friday	Morning	1	77
	Tuesday	Morning	1	83
	Wednesday	Morning	1	87
Ushuluddin Faculty	Tuesday	Morning	1	74
	Wednesday	Morning	1	74
Library	Monday	Morning	1	91
	Monday	Afternoon	1	85
	Wednesday	Afternoon	2	185
	Wednesday	Evening	1	78
Convention Hall	Tuesday	Morning	1	144
	Wednesday	Morning	1	76

VI. CONCLUSSION

The number of internet users at the Sunan Kalijaga State Islamic University through the WiFi SUKAnet network is quite a lot. From the analysis conducted by the researcher, the following results are obtained:

1. The highest number of users in a week occurs in the morning.
2. The number of users in a week from morning to evening has a pattern that always goes down.
3. There are 13 access point points that have very high access loads in a week.
4. There are four locations that have very high access load values, namely the Adab and Cultural Sciences Faculty, the

Science and Technology Faculty, the Ushuluddin Faculty, the Library.

5. Of the four locations only in the Library, the access point is consistent with a very high access load with an average access point load value of 91 users.

REFERENCES

[1] Arsih, Kansul. (2012). *Perancangan dan Implementasi Aplikasi Analisis Log Menggunakan Metode JST Adaptive Resonance Theory 2 dalam Memprediksi Tingkah Laku Pengguna Internet*. Bandung : Universitas Telkom

[2] Witten, et al. (2012). *Data Mining Practical Machine Learning Tools and Technique, 2nd Edition.* San Fransisco : Morgan Kaufmann

[3] Micro, Andi. (2012). *Dasar-dasar Jaringan Komputer*. clearOS Indonesia

[4] Aini, Fithratul. (2011). *Web Usage Mining Menggunakan Algoritma Adaptive Web Access Pattern Tree (AWAPT)*. Bandung: Universitas Telkom

[5] Han, Jiawei. et al. (2011). *Data Mining: Concepts and Tecniques, 3rd ed.* San Francisco: Morgan Kauffman

[6] Prasetyo, Eko. (2012). *Data Mining : Konsep dan Aplikasi Menggunakan MATLAB.* Yogyakarta : Andi