Journal of Theoretical and Applied Information Technology
20th January 2016. Vol.83. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DENIAL OF SERVICE LOG ANALYSIS USING DENSITY K-MEANS METHOD

[1]ARDYMULYA ISWARDANI, [2]IMAM RIADI

[1]Department of Informatics, Indonesia Islamic University, Yogyakarta

[2]Department of Information Systems, Ahmad Dahlan University, Yogyakarta

E-mail: [1]am.iswardani@gmail.com, [2]imam.riadi@mti.uad.ac.id

ABSTRACT

Denial of service attacks launched by flooding the data on the ftp server causes the server to be unable to handle requests from legitimate users, one of the techniques in detecting these attacks is by monitoring, but found several problems including the difficulty in distinguishing the attack and with normal data traffic. So that the necessary field studies of triage forensics to get a vital information at the scene that is useful in supporting the overall digital forensics investigation. Triage forensics begins with the log databases which are then performed by using the grouping density k-means algorithm to produce three levels of danger (low, medium and high).

Proposed density k-means algorithm using three groups that represent the level of danger. The minimum value, medium, and maximum of the dataset as early centroid, the data which has minimum distance to the centroid value specified will join to form a cluster centroid. Data that has been joined in a cluster and then evaluated the level of density (density) with its center (centroid) using Davies-Bouldin index.

Results of clustering that has been done in the dataset resulted in three clusters, but the level of danger which successfully identified only two, namely the level of danger of medium and high, the value of DBI obtained 0.082, indicates that the data used manifold homogeneous, results DBI obtained is also influenced by the selection of the value of the centroid beginning clustering process.

Keywords: *Clustering, Triage Forensic, Log, Analysis, Density K-Means*

1. INTRODUCTION

Motivation in this study originated from reports[1], related attacks that often occur in indonesia, The focus of this research is a type of DoS attack that has a way of working by sending mass demand on the server until unable to serve the demand[2]. Users often do not realize that the system may be a target[3] it is due to the difficulty in distinguishing normal and abnormal traffics.

In order to recognize the DoS attack can be done by monitoring network traffic, which aims to recognize the existence of an attack but it still encountered some obstacles[4]. Based on the above explanation, the study of forensic triage is needed to extract the vital information on site for supporting the investigation matter[5].

Digital forensics is a methodology that relates to the recovery process and the investigation material found on digital evidence, as part of the investigation[6] so that the facts found in digital evidence acceptable in court. The relation between forensic triage with digital forensics is on its findings are used to support the digital forensic process.

Forensics triage process begins from the input of database logs obtained during the monitoring process then the database is grouped with the clustering using k-means algorithm which aims to classify the danger level of low, medium and high. clustering utilized to help manage the complexity in managing large databases.

Clustering as the techniques used in classifying the data[7]. Clustering will split the data into several groups according to the specific characteristics based on the calculation of the distance data more closely than the other data[8], [9]. In statistics and machine learning, k-means algorithms is a cluster analysis method that leads to the observation object of partitioning N into K where each observation object is owned by a group this research proposed Density k-means algorithm to classified the data to produce three level danger. The data has been incorporated in the cluster formed is then determined how close (density) of data with its centroid.

Log came from network traffic serves to identify the presence or absence of a DoS attack. Logs are stored in the original format in the form of text and then stored in a database. Log has a large size, therefore it is necessary to do some measures to ease the process of storage and retrieval of information in the database.

2. BASIC THEORY

2.1. K-means Algorithm

K-means is one method of non-hierarchical clustering of data that seek to partition the existing data in the form of one or more clusters. This method is partitioned into clusters so that the data that have the same characteristics are grouped into the same cluster and that has different characteristics grouped other clusters[10], K-means algorithm is show in figure 2.1.

1. Initialization: Determine the value of *K* as the number of clusters desired.
2. Select the *K* data from the dataset as a centroid
3. Allocate all data to the nearest centroid by a predetermined distance metric.
4. Recalculate centroid *C* based on data that follows each cluster.
5. Repeat steps 3 and 4 until the convergent condition is reached (no data is moved).

FIGURE 2.1: K-MEANS ALGORITHM

2.2. Distance

To measure the non-resemblance the two data with multiple attributes for each quantity of data used distance (distance). There are many models of distance measurement, and the most commonly used is the euclidean distance[9].

$$D(\chi, y) = \|\chi - y\|_2 = \sqrt{\sum_{j=1}^{N} |x - y|^2} \qquad (1)$$

Where *D* is the distance between *x* and *y*, *n* is the absolute value. *N* is the number of features (dimension) data.

2.3. Davies-Bouldin Index (*DBI*)

Method of doing an internal evaluation based cluster partition use Davies-Bouldin Index which has characteristics in validating clusters based on calculations derived from the quantity and feature datasets

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{i \neq j} \left(R_{i,j} \right) \qquad (2)$$

Where *DBI* is obtained scalar value, *K* is number of clusters used. Essentially, DBI want value as small as possible to assess the cluster obtained good[9].

3. METHODOLOGY

Clustering with density k-means on this study illustrated in figure 3.1.

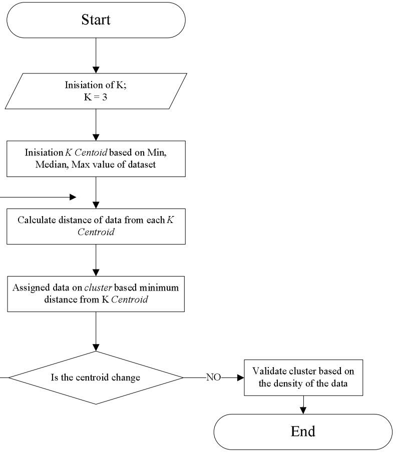

Start
Inisiation of K;
K = 3
Inisiation *K Centoid* based on Min, Median, Max value of dataset
Calculate distance of data from each *K Centroid*
Assigned data on *cluster* based minimum distance from K *Centroid*
Is the centroid change
YES
NO
Validate cluster based on the density of the data
End

FIGURE 3.1: DENSITY K-MEANS ALGORITHM

Explanation of figure 3.1. as follows:

1. Initiate *k*, where *k* is the number of clusters (groups) to be formed.

2. Determining the initial centroid obtained from the minimum, medium and maximum value of dataset.

3. Calculate the distance of each data with each centroid.

4. Group data on a cluster based on the minimum distance of the centroid.

5. When the value *newCentroid* with *oldCentroid* does not change, then the clustering process by using algorithms Density k-means finished, but when the value newCentroid with *oldCentroid* still changing, repeat the process of calculating the distance to a value not changed.

f. The results of the last iteration clustering process then validated using internal validation techniques Davies-Bouldin Index.

4. RESULT

4.1. Clustering Result

In this study it is determined that the level of danger can be evaluated based on the smallest value, the middle value and the greatest value of total second attribute. Classification of hazard levels shown in Table 4.1.

Table 4.1. Danger Level Criteria

	Danger Levels	Tot length	Tot tcplength
Min	Low	2619	0
Median	Medium	3685.5	2
Max	High	10104895	4971517

After the clustering process is done, the results are shown by table 4.2. Known of labeling are two

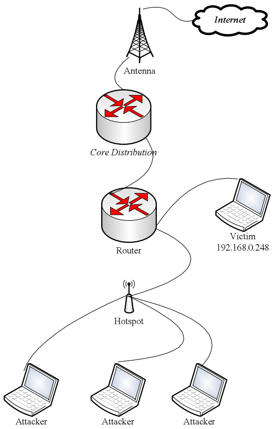
(2) types of hazard levels are medium and high. This labeling result is influenced by the quality of the data that is homogeneous and the effect of early elections centroid value. Clustering process that has been completed in two (2) data features: length and tcplength database *log* as much as 11.358.001, in Table 4.2 cluster to the level of danger of being a membership of 22, the data is data 1st until the 16th of data, the data of the 18th to the 23rd of data; then cluster with the level of danger was a membership of one member of that data to the 17th; and clusters with high danger level has a membership of one member of that data to the 24th. Then for it is supported by *DBI* value obtained by 0.082870025, which means the structure, membership and compactness of data which are members of each cluster-value (distance) 0.

Table 4.2 Danger Level

Hour-*i*	***length***	***tcplength***	***Classification***
1	3404	2	MEDIUM
2	2619	1	MEDIUM
3	3862	1	MEDIUM
4	2948	0	MEDIUM
5	3871	1	MEDIUM
6	4622	0	MEDIUM
7	3163	0	MEDIUM
8	3253	2	MEDIUM
9	2911	1	MEDIUM
10	2648	0	MEDIUM
11	3376	12	MEDIUM
12	2692	0	MEDIUM
13	3020	0	MEDIUM
14	3658	0	MEDIUM
15	9809	2688	MEDIUM
16	26871	6905	MEDIUM
17	983613	355522	MEDIUM
18	3973	100	MEDIUM
19	3713	84	MEDIUM
20	4114	90	MEDIUM
21	4029	69	MEDIUM
22	3416	84	MEDIUM
23	160322	85971	MEDIUM
24	10104895	4971517	HIGH

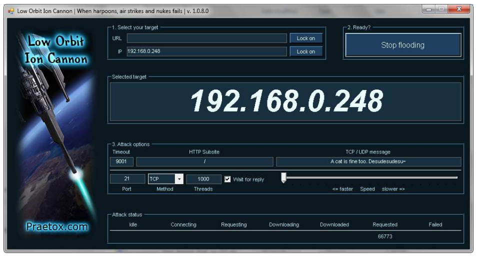
4.2. Attack Simulation

Attack simulation illustrated by figure 4.1. This research applies forensic triage framework which then conducted an analysis of database logs that have been obtained. Tests conducted on the local network of SMEs Mandala Citra Media in Surakarta, where an attacker use the tool LOIC (Low Orbit Ion Canon) and FTP BruteForce. Test scenario shown in Figure 4.1. explained that the victim has the IP address 192.168.0.248 connected to the network through a router. In addition there are some attacker (attacker) that pass through the local network attacks.



FIGURES 4.1. ATTACK SIMULATION

a. LOIC


Low Orbit Ion Cannon | When harpoons, air strikes and nukes fails | v. 1.0.8.0

1. Select your target
URL
IP 192.168.0.248

Selected target
192.168.0.248

3. Attack options
Timeout 9001
HTTP Subsite /
TCP / UDP message A cat is fine too. Desudesudesu~
Port 21
Method TCP
Threads 1000
☑ Wait for reply

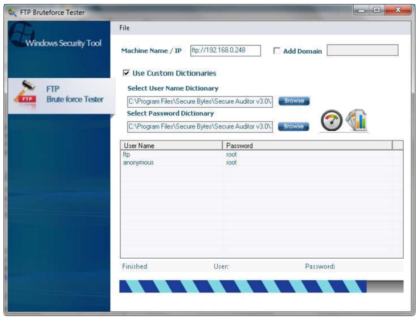
Attack status
Requested 66773

FIGURE 4.2: LOIC

The advantage of this tool can perform packet delivery request based protocols tcp or udp. Besides the target port to be sent can be specified by the attacker. Simulation tool in a denial of service shown in the figure 4.2.

b. FTP BruteForce

A tool that uses brute force method to obtain user information such as username and password to access the ftp server, the advantages of this tool because of its ability in recognizing weaknesses that are owned by the ftp server, this tool works by testing all combinations of usernames and passwords are commonly used. Simulation tool in a denial of service shown in the figure 4.3.



Machine Name / IP: ftp://192.168.0.248
□ Add Domain
☑ Use Custom Dictionaries
Select User Name Dictionary: C:\Program Files\Secure Bytes\Secure Auditor v3.0\
Select Password Dictionary: C:\Program Files\Secure Bytes\Secure Auditor v3.0\

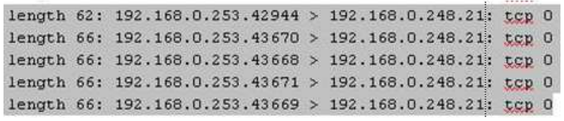
User Name	Password
ftp	root
anonymous	root

Finished User: Password:

Figure 4.3: Ftp Bruteforce

4.3. Log Verification

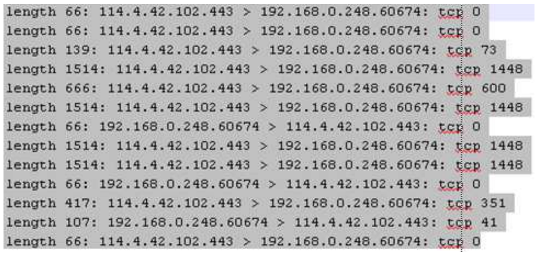
Based on the results that have been obtained from the victim machine, the next step is to verify the information from the log database analysis with the original logs that are stored in text files. For DoS attacks on port 21 with LOIC can know the result that there has been an attack on the system through port 21 to the length of the header 66 as shown in Figure 4.4.



```
length 62: 192.168.0.253.42944 > 192.168.0.248.21: tcp 0
length 66: 192.168.0.253.43670 > 192.168.0.248.21: tcp 0
length 66: 192.168.0.253.43668 > 192.168.0.248.21: tcp 0
length 66: 192.168.0.253.43671 > 192.168.0.248.21: tcp 0
length 66: 192.168.0.253.43669 > 192.168.0.248.21: tcp 0
```

Figure 4.4: Log Verification(A)

Based on the results that have been obtained from the victim machine, the next step is to verify the information from the log database analysis with the original logs that are stored in text files. For DoS attacks with FTP Bruteforce can know the results that have been attacks on the system through port 443 to the length of the header 66 and tcplength 0 as shown in figure 4.5



```
length 66: 114.4.42.102.443 > 192.168.0.248.60674: tcp 0
length 66: 114.4.42.102.443 > 192.168.0.248.60674: tcp 0
length 139: 114.4.42.102.443 > 192.168.0.248.60674: tcp 73
length 1514: 114.4.42.102.443 > 192.168.0.248.60674: tcp 1448
length 666: 114.4.42.102.443 > 192.168.0.248.60674: tcp 600
length 1514: 114.4.42.102.443 > 192.168.0.248.60674: tcp 1448
length 66: 192.168.0.248.60674 > 114.4.42.102.443: tcp 0
length 1514: 114.4.42.102.443 > 192.168.0.248.60674: tcp 1448
length 1514: 114.4.42.102.443 > 192.168.0.248.60674: tcp 1448
length 66: 192.168.0.248.60674 > 114.4.42.102.443: tcp 0
length 417: 114.4.42.102.443 > 192.168.0.248.60674: tcp 351
length 107: 192.168.0.248.60674 > 114.4.42.102.443: tcp 41
length 66: 114.4.42.102.443 > 192.168.0.248.60674: tcp 0
```

Figure 4.5: Log Verification(B)

5. CONCLUSION

Algorithm Density K-means can be used to help perform grouping database log is based on the danger level is low, medium or high, although the result was only found two types of degree of danger is the danger of medium and high, however the algorithm density k-means capable of partitioning the data into 3 (three) cluster group.

Application of forensic triage framework can be applied in making the initial identification to recognize the existence of a DoS attack, based on the stage: a) the classification of electronic devices; b) perform the extraction of digital devices; c) perform feature extraction as an input in the analysis process; d) processing algorithms in the analysis used in this case is the density k-means; and e) reporting of findings that serves as a guide to support the existence of an incident to immediately take steps (decide) early remedial action to reduce the level of damage to the system.

REFRENCES:

[1] ID-SIRTII/CC, “Data Internet Trafik Tahun 2014,” *ID-SIRTII/CC*, 2014. [Online]. Available: http://www.idsirtii.or.id/tahunan/tahun/2014.html. [Accessed: 10-Jul-2015].

[2] A. Kak, “TCP/IP Vulnerabilities: IP Spoofing and Denial-of-Service Attacks,” pp. 1–103, 2015.

[3] A. Kurniawan, *Network Forensics Panduan Analisis & Investigasi Paket Data Jaringan*. Yogyakarta: Penerbit ANDI, 2012.

[4] Z. Tan, A. Jamdagni, X. He, P. Nanda, and R. P. Liu, “A system for denial-of-service attack detection based on multivariate correlation analysis,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 2, pp. 447–456, 2014.

[5] D. Mcclelland and F. Marturana, “A Digital Forensics Triage Methodology based on Feature Manipulation Techniques,” in *Communications Workshops (ICC* ., 2014, pp. 676 – 681.

[6] I. Riadi, J. E. Istiyanto, A. Ashari, and Subanar, “Internet Forensics Framework Based-on Clustering,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 12, pp. 115–123, 2013.

[7] S. H. Liao, P. H. Chu, and P. Y. Hsiao, “Data mining techniques and applications - A decade review from 2000 to 2011,” *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11303–11311, 2012.

[8] E. Prasetyo, *Data Mining Konsep dan Aplikasi Menggunakan Matlab*. ANDI Offset, 2012.

[9] E. Prasetyo, *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab*. ANDI Offset, 2014.

[10] Narwati, “Pengelompokan Mahasiswa Menggunakan Algoritma K-means,” *J. Din. Inform.*, vol. 2, no. 2, 2010.