ANALYSIS EMAIL CONTENT WITH K-MEANS CLUSTERING FOR PROFILING ON POSTFIX SERVER

[1]BAMBANG SUGIANTORO, [2]MUSLIM HERI KISWANTO

[1], Department of Informatics Engineering, Sunan Kalijaga State Islamic University, Yogyakarta, Indonesia

[2]Department of Informatics Engineering, University of Duta Bangsa (STMIK), Surakarta, Indonesia

E-mail: [1]Bambang.sugiantoro@uin-suka.ac.id, [2]heri_muskis@yahoo.co.id,

ABSTRACT

Clustering email content is a way of categorization of email content based on specific criteria to obtain the category types of emails that can be used for content profiling email. There are several types and content of the email on a mail server that is not desired by the client so that if this occurs continuously will result in performance mail server will be disrupted. Email content profiling was conducted to determine patterns and behavior in the Postfix SMTP email delivery server using the K-means algorithm to know the type of email that can disrupt the performance of the mail server. The profiling process begins with taking the log file content stored on the Postfix Simple Mail Transfer Protocol (SMTP) email server and then analyze the log file using Clustering techniques with K-means algorithm for Profiling email content on Postfix-based mail server using three criteria, namely: message_id, access from and content. Email content analysis was conducted using K-Means Clustering until the 3rd iteration. The clustering results obtained from four categories of e-mail stored in the postfix mail server, that is a true email, fraud, advertising, and emails which do not have a clear purpose. The step profiling using the deductive method to the content of the email that will be obtained information about the type and characteristics of e-mail stored in the Postfix SMTP server. Using the deductive method for the profiling is the solution to assign profiles based on data obtained from the scene, in this case, is the content of the email on the SMTP Mail Server. There are 3 main things done in this deductive technique, namely: analysis of the operation mode refers to the clustering process, as a characteristic signature in email delivery and victimology identified as a victim in email delivery.

Keywords: *Profiling, Email, K-Means clustering, Postfix*

1. INTRODUCTION

Internet and email have changed the world with annual growth increasing with the development of mobile technology and the email itself [1] [2]. Then from an existing email content, nearly a quarter are expected to e-mail spam or not accepted by the account owner. The rise of cybercrime today also an impact on e-mail users, it also can be seen the number of distribution messages or unwanted data example is junk email [3]. Spam email is one of the most complex problems in the mail service [4] [5]. Email spam is unsolicited email, which is not addressed to specific recipients delivered whether for marketing purposes, or for frauds, and hoaxes [6].

In some cases, the abuse of email, the sender will try to hide real identity to avoid detection. For example, a false return address or anonymized by routing e-mail through the mail server anonymously, or e-mail content and header information which has been modified to conceal the identity of the sender [7][8]. To find out what content is on the email server for delivery of traffic to occur, the need for categorization of the content [9]. In conducting the necessary categorization data mining technique with clustering algorithms aimed at grouping several data/objects into groups/clusters and create distance between clusters as far as possible [10].

The amount incoming email to the mail server could lead to disruption of mail service to the client, especially if an incoming email is largely a kind of email that could interfere with the performance of mail servers as spam and junk mail. So, it needs mapping and classification of content on the mail server to determine the type of content is useful and junk mail. By using clustering techniques in the categorization of the type of content on Postfix Simple Mail Transfer Protocol (SMTP) server can be used as a solution [11]. Clustering types of content that no mail server can assist in optimizing mailing services, because it will know what is good or link content objects that are often sent to the mail server.

In this papers in our research on email content analysis used algorithm Clustering with K-Means. Profiling for the process on the stages Postfix SMTP server is profiling email content with K-Means Cluster on Postfix SMTP Server. The first stage begins by capturing the log file from the SMTP server, and then analyzed K-Means Clustering using a tool RapidMiner, the next stage is profiling the content of the email Profiling methods used in this study is the deductive method, namely setting up profiles based on data obtained from the scene, in this case, is the content of the email on the SMTP Mail Server.

The results we present in this paper is a K-Means Clustering Algorithm can be used to analyze the types of email content on postfix mail server to obtain four categories of email, namely: True email, Scam / Fraud, Promotion and Unpurpose, by classifying each pattern characteristic of mail by 3 criteria, namely message_id, Access from and Content. Profiling email content with K-Means Clustering method can classify types of email content in accordance with the character, so it can be used to filter the content of the category of unsolicited email, promotion, fraud and performance of Postfix SMTP Mail Server work optimally

2. BASIC THEORY

2.1 Data Mining

Data mining is an interactive and interactive process to find patterns or new models are valid, useful and can be understood into a very large database [12] [13]. Data mining provides the search for patterns or trends are desirable in large databases to assist decision making in the future[8]. Where these patterns recognized by a device that can provide a useful analysis and insightful data that can then be studied more closely, which may use other decision support tools [3].

Data mining commonly involves four classes of tasks:

a. Clustering: groups and structures in the data
b. Classification: generalizing known structure to apply to new data. For example, an email program might attempt to classify an email is legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification, neural networks, and support vector machines.
c. Regression: the effort to find a function which models the data with minimal errors.
d. Association rule learning: the search for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning as market basket analysis, for example, the supermarket can determine which products are frequently bought together and use this information for marketing purposes[4].

2.2 Clustering

Clustering is one of data mining technique that is often used with the aim to identify homogeneous groups or clusters of a set of objects [5] [6]. This technique goes by partitioning the data set into several subsets or groups such that the elements of a particular group have a set of properties that are shared, with a high degree of similarity in one group and the degree of similarity between the groups is low. Clustering process is to divide activity data in a set into groups of data commonality within a group is greater than the similarities that data with data in other groups [7]. Shown in Figure 1 Different ways of clustering the same set of points[8]

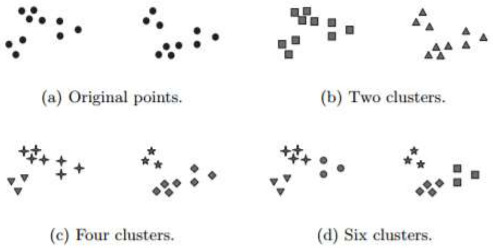


Figure 1: Different ways of clustering the same set of points

2.3 Algoritma K-Means

K-means clustering is a method of non-hierarchical clustering of data that categorize the data in the form of one or more clusters/groups [6]. Data that has the same characteristics are grouped into one cluster/group and data that have different characteristics grouped by cluster/group to another so the data are in one cluster/group has a small degree of variation [14] [15] [16]

This algorithm is widely used for partition-based clustering, each point is inserted at the center point determined by the distance calculation, K-Means algorithm to perform clustering will do looping until the center point unchanged [16]. In our study is the value of X determined in four categories of email content to be inserted shown in figure 2.

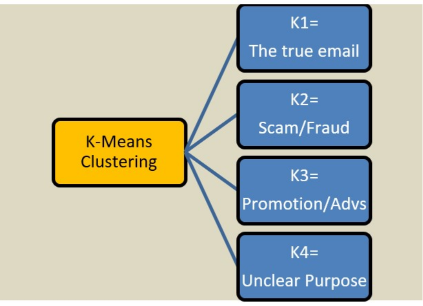


Figure 2: Clustering K-Means algorithm

Explanation of figure 2 is:

a. The true email, is a useful email, meaning that the account owner desires email.

b. Scam / Fraud, an email which aims to commit fraud, of this type can be:
   1. Lottery/award
   2. Business proposals
   3. Requests for favors
   4. Financial offers (e.g. loans or credit cards)
   5. Phishing

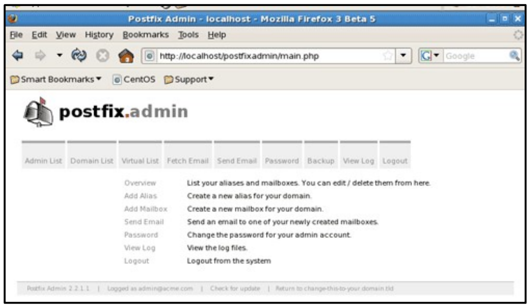
c. Promotion/advertising, an email that aims at offering certain products that are not specified by the user.

d. Unclear Purpose, an email content that has the purpose unclear

2.4 Postfix Mail Server

Postfix is software MTA (Mail Transfer Agent) that can receive, forward, and send an email. This a software is open source and works on Unix and Linux OS [17], postfix does not support windows. Postfix installation generally performed to replace send mail program by default has been installed on the operating system Unix or Linux [18]. Postfix can be an option for mail server software for the claimed performance and high speed, besides Postfix also supports the use of the database [18]. Beginning of Postfix named V mailers and IBM Secure Mailer, but for reasons of similarity with existing trademark, then changed to postfix on IBM's own suggestion. Wietse Venema is the creator postfix when he worked at IBM, and until now still being developed.

In terms of security and modularity postfix positioning itself as a rival mail server artificial D. J. Bernstein, both are very concerned about security and modularity. Various people argue about Postfix faster. Postfix is implementing a reliable alternative to Sendmail program has been widely used in UNIX sites. Postfix can do things such as virtual domains, which have multiple domains on the same computer, the Control Host to blacklist certain hosts and many others


Postfix Admin - localhost - Mozilla Firefox 3 Beta 5
File Edit View History Bookmarks Tools Help
http://localhost/postfixadmin/main.php
Google
Smart Bookmarks
CentOS
Support
postfix.admin
Admin List
Domain List
Virtual List
Fetch Email
Send Email
Password
Backup
View Log
Logout

Overview	List your aliases and mailboxes. You can edit / delete them from here.
Add Alias	Create a new alias for your domain.
Add Mailbox	Create a new mailbox for your domain.
Send Email	Send an email to one of your newly created mailboxes.
Password	Change the password for your admin account.
View Log	View the log files.
Logout	Logout from the system

Postfix Admin 2.2.1.1 | Logged as admin@acme.com | Check for update | Return to change-this-to-your domain.tld

Figure 3: Postfix Mail server

Some excellent features of Postfix as a reliable email system [18]:

a. Multi-transport: Postfix designed flexible enough where he can operate in a variety of environments, such as Internet, DECnet, and UUCP, without requiring a virtual domain. Even so, the initial release of Postfix admittedly can only communicate with SMTP and limited to UUCP.

b. Virtual Domains: Adding virtual domains in Postfix is easy enough where we just need to change a single lookup table, while the other of mailers generally require multilevel aliasing or redirect to obtain the same results.

c. Restriction Relay: Postfix provides a way for us to host restriction, the name of which can relay mail through Postfix system, and which one is permitted to enter the mail. For Postfix needs to implement the operation blacklist, RBL lookups, HELO/sender DNS lookups.

d. Table lookups: Postfix does not implement address rewriting language, it employs what is called a table lookups. The tables can be local DB or DB files, or other lookup mechanisms are also quite easy.

3. METHODOLOGY

3.1 Network Topology Scenarios

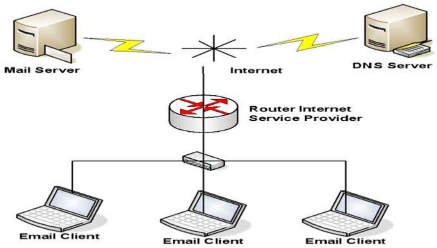


Figure 4: Network Topology Scenarios

Following is an explanation of each component contained in the network topology in Figure 4:

a. DNS Server is used to translate a domain name into an IP address used by the Public Mail Server or otherwise. Design scenarios mail server is registered with the primary domain webmail.herismadelta.id
b. Mail Server, designed as the main server where taking research data. In this design, Mail Server will be set up in a VPS (Virtual Private Server) with Public IP 219.83.63.150 with the following specifications: 1.7 GHz Quad Core Processor, 1 GB Ram, 40 GB hard drive, network card 2 virtual mode, system Debian 6.0.1 operation Systems using Postfix Mail Server, POP3, IMAP4 and web Mail
c. The Internet is a large-scale network that connects computers to the Mail Server client through an ISP that allows users to access the mail service.
d. Router Internet Service Providers, Hardware which serves as a gateway bridging communication with the client computer services through the Internet Mail Server.
e. Email Client is a computer used by clients in the use of the mail service. Applications that are used can be varied, ranging applications such as web browsers Chrome and Mozilla.

3.2 Log Capturing

The capture process log files of server systems were conducted from September 1 to December 1, 2016, with data samples taken on mail originating access service network High School, where the mail service users are students. A log file that will capture and parsed as follows: /var/log/auth.log, / var / log / lastlog, /var/log/mail.log, /var/log/mail.info, / var / log /user.log, /var/log/mail.warn and /var/log/apache2/access.log.

3.3 Clustering Mail Content

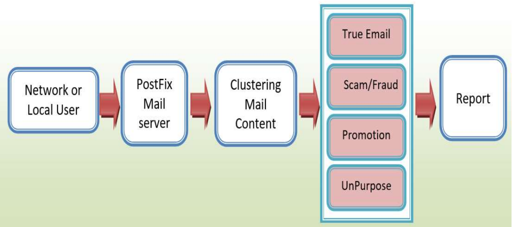


Figure 5 Flow Stages Clustering Mail Content

Explanation of Figure 5 on Clustering Mail Flow Stages Content is the first stage in making an email sent clustering is to describe the area of network and email users who have registered in the system. Then the second phase is to create a policy rule in the Postfix queue so that every email that is sent is always recorded in the log file. The third stage is to force every email that is sent to the recorded content and headers always be a base64 encryption-based file. The fourth step is to perform clustering the content of the email, and the last step is to write a report.

Clustering process based on the number of emails that go out or go to the postfix mail server clustered by type into four categories, namely:

a. Email True, symbolized by **T**
b. Scam / Fraud is fraud email, symbolized by **S**
c. Promotion is the email form of advertising / promotional event of a product, symbolized by **P**
d. Unpurpose that email has no clear purpose or misdirected, symbolized by **U**

Determine the weight of each criterion cluster email content. The criteria used in conducting email cluster consisting of:

a. Message ID

Email sent via postfix mail server will be recorded and analyzed based on its address. Table 3.1 explains the weight of variable message_id:

Tabel 3.1 Weight criteria message_id

information	Weight	Analysis Email Address	No
High	8-9	Email client IDXXXX postfix form (example ID00212)	1
Medium	5-8	Webmail (Gmail dan yahoo)	2
Low	1-5	Email that is rarely used or online stores (eg revasia.com, lazada)	3
Empty	0	no email address (null)	4

b. Content

The analysis is done based on email content every word (string), which has synonymous with the division of the cluster. Table 3.2 explains the weight of variable Content:

Tabel 3.2 Weight criteria content

information	Weight	Analysis Content	No
High	8-9	There is a string associated with the school, lessons, and introductions in Indonesian	1
Medium	5-8	There is a string associated with product offerings	2
Low	1-5	There is a string with foreign terms (especially English)	3
Empty	0	no content	4

c. Access from

This analysis is based on access to email delivery, which generally comes from HTTP, IMAP, telnet or crontab. Table 3.3 explains the weighting of variables Access from:

Tabel 3.3 Weight Criteria Access from

information	Weight	Access Analysis From	No
High	8-9	Imap	1
Medium	5-8	HTTP	2
Low	1-5	telnet dan crontab	3
Empty	0	null	4

Preliminary data clustering process testing for K-Means algorithm shown in Table 3.4:

amount	access form	content	message_id	email
26	8	9	9	**1**
21	7	6	8	**2**
8	2	2	4	**3**
8	2	3	3	**4**
24	9	8	7	**5**
4	1	0	3	**6**
4	0	0	4	**7**
24	8	8	8	**8**
26	9	8	9	**9**
14	2	3	9	**10**
12	2	5	5	11
24	8	9	7	12
24	8	8	8	13
2	0	2	0	14
11	2	4	5	15
21	8	4	9	16
27	9	9	9	17
12	2	5	5	18
4	0	4	0	19
25	9	8	8	20

The six stages of the process of clustering mail content, namely:

1. Determine the value of k as the number of clusters to be formed.

Determining 4 cluster data representing the type of email, namely: True mail, Scam / Fraud, Promotion, and UnPurpose. Data is assumed to be:

1. Data to 1 as cluster-1 (9,9,8)
2. Data to 3 as cluster-2 (4,2,2)
3. Data to 6 as cluster 3rd (3,0,1)
4. Data to 10 as a cluster 4th (9,3,2)

2. Initialize K centroid.
   Initialization determines the starting position of each cluster centroid is done by using a value that indicates the proximity with the category. Example message_id premises mail-n value = 9 and the access from the value = 4 and Content to a value of 2, three grades of both attributes are used as initialization centroid
3. Calculate the distance.
   The process of calculating the distance to the centroid similarity to the email content C1 using equation 3.1

$$(\;,\;) = |\;\; - \;\;| = \sqrt{\sum_{=1} |\;\; - \;\;|^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(3.1)$$

Centroid to1:

$$= \sqrt{(9-9)^2 + (9-9)^2 + (8-9)^2}$$

$$= \sqrt{0+0+1}$$

$$= \sqrt{1}$$

= 1

Centroid to 2:

$$= \sqrt{(9-4)^2 + (9-2)^2 + (8-2)^2}$$

$$= \sqrt{5^2 + 7^2 + 6^2}$$

$$= \sqrt{25+49+36}$$

$$= \sqrt{110}$$

= 10.5

Centroid ke-3:

$$= \sqrt{(9-3)^2 + (9-0)^2 + (8-1)^2}$$

$$= \sqrt{6^2 + 9^2 + 7^2}$$

$$= \sqrt{36+81+49}$$

$$= \sqrt{166}$$

= 12.9

Centroid to 4

$$= \sqrt{(9-9)^2 + (9-3)^2 + (8-2)^2}$$

$$=\sqrt{0^2\ + 6^2 + 6^2}$$
$$= \sqrt{0 + 36 + 3649}$$
$$= \sqrt{72}$$

= 8.5

The following is shown in Table 3.5 is a calculation Iteration 1:

Tabel 3.5 Calculation Iteration 1

Shortest distance	cent 4	cent 3	cent 2	cent 1	sum	access form	content	messag e_id	email
	2 3 9	1 0 3	2 2 4	9 9 9					
1.00	8.49	12.88	10.49	1.00	26	8	9	9	1
3.74	5.92	9.85	7.55	3.74	21	7	6	8	2
0.00	5.10	2.45	0.00	11.09	8	2	2	4	3
1.41	6.00	3.16	1.41	11.00	8	2	3	3	4
2.24	8.83	12.00	9.70	2.24	24	9	8	7	5
0.00	6.78	0.00	2.45	13.45	4	1	0	3	6
1.41	6.16	1.41	2.83	13.67	4	0	0	4	7
1.73	7.87	11.75	9.38	1.73	24	8	8	8	8
1.00	8.60	12.81	10.49	1.00	26	9	8	9	9
0.00	0.00	6.78	5.10	9.22	14	2	3	9	10
3.16	4.47	5.48	3.16	9.00	12	2	5	5	11
2.24	8.72	12.08	9.70	2.24	24	8	9	7	12
1.73	7.87	11.75	9.38	1.73	24	8	8	8	13
3.74	9.27	3.74	4.47	14.53	2	0	2	0	14
2.24	4.12	4.58	2.24	9.49	11	2	4	5	15
5.10	6.08	10.05	8.06	5.10	21	8	4	9	16
0.00	9.22	13.45	11.09	0.00	27	9	9	9	17
3.16	4.47	5.48	3.16	9.00	12	2	5	5	18
4.90	9.27	5.10	4.90	13.67	4	0	4	0	19
1.41	8.66	12.37	10.05	1.41	25	9	8	8	20

4. Grouping Data

Distance on the calculation will be performed and selected comparison data with the shortest distance between the center of the cluster, this range indicates that the data are in one group with the nearest cluster center. Shown in Table 3.6 grouping matrix data group, the value of 1 means that the data is in the group

Tabel 3.6 Grouping Data in Iteration 1

Cen 4	Cen 3	Cen 2	Cen 1	Email
0	0	0	1	**1**
0	0	0	1	2
0	0	1	0	**3**
0	0	1	0	4
0	0	0	1	5
0	1	0	0	**6**
0	1	0	0	7
0	0	0	1	8
0	0	0	1	9
1	0	0	0	**10**
0	0	1	0	11
0	0	0	1	12
0	0	0	1	13
0	1	0	0	14
0	0	1	0	15
0	0	0	1	16
0	0	0	1	17
0	0	1	0	18
0	0	1	0	19
0	0	0	1	20

5. Determination of the new cluster

Once known members of each cluster and then new cluster centers are calculated based on data from each cluster member in accordance with the formula of the center cluster members. Obtained calculation of each new cluster is shown in Table 3.7

Tabel 3.7 Iteration Calculation To 2

Shortest distance	cent 4 NEW	cent 3 NEW	cent 2 NEW	cent 1 NEW	Sum	access form	content	message_id	email
	2 3 9	5 0 2	11 2 4	8 8 8					
1.56	8.49	11.63	9.25	1.56	26	8	9	9	1
2.15	5.92	8.58	7.03	2.15	21	7	6	8	2
3.73	5.10	3.73	8.84	9.48	8	2	2	4	3
3.90	6.00	3.90	8.91	9.42	8	2	3	3	4
1.42	8.83	10.13	7.10	1.42	24	9	8	7	5
3.82	6.78	3.82	10.06	11.82	4	1	0	3	6
5.09	6.16	5.09	11.02	12.08	4	0	0	4	7
0.47	7.87	10.29	7.92	0.47	24	8	8	8	8
1.10	8.60	11.25	8.23	1.10	26	9	8	9	9
0.00	0.00	7.95	10.37	7.90	14	2	3	9	10
4.47	4.47	6.16	9.42	7.56	12	2	5	5	11
1.79	8.72	10.55	8.25	1.79	24	8	9	7	12
0.47	7.87	10.29	7.92	0.47	24	8	8	8	13
5.34	9.27	5.34	11.44	12.99	2	0	2	0	14
4.12	4.12	5.44	9.15	7.98	11	2	4	5	15
3.80	6.08	8.58	6.36	3.80	21	8	4	9	16
1.68	9.22	11.95	8.99	1.68	27	9	9	9	17
4.47	4.47	6.16	9.42	7.56	12	2	5	5	18
6.26	9.27	6.26	11.61	12.24	4	0	4	0	19
0.79	8.66	10.66	7.62	0.79	25	9	8	8	20

6. Repeat steps 3rd (three) if the data is still changing.

Distance on the calculation will be performed and selected comparison data with the shortest distance between the center of the cluster, this range indicates that the data are in one group with the nearest cluster center. The following are shown in Table 3.8 grouping matrix data group, the value of 1 means that the data is in one group, the following table illustrates the group for grouping data iteration 2:

Tabel 3.8 Iteration Data Grouping to 2

Cen 4	Cen 3	Cen 2	Cen 1	email
0	0	0	1	1
0	0	0	1	2
0	1	0	0	3
0	1	0	0	4
0	0	0	1	5
0	1	0	0	6
0	1	0	0	7
0	0	0	1	8
0	0	0	1	9
1	0	0	0	10
1	0	0	0	11
0	0	0	1	12
0	0	0	1	13
0	1	0	0	14
1	0	0	0	15
0	0	0	1	16
0	0	0	1	17
1	0	0	0	18
0	1	0	0	19
0	0	0	1	20

Since the result of grouping data iteration 1 iteration 2 are not the same and it is necessary to iteration 3 as shown in Table 3.9, the same way as step 2, for more details see the results table 3rd iteration of the following:

Tabel 3.9 Iteration Calculation To 3

Shortest distance	cent 4 NEW	cent 3 NEW	cent 2 NEW	cent 1 NEW	sum	access form	content	message_id	email
	2 4 6	1 2 2	0 0 0	8 8 8					
1.56	8.22	12.13	15.03	1.56	26	8	9	9	1
2.15	5.66	9.35	12.21	2.15	21	7	6	8	2
2.04	3.01	2.04	4.90	9.48	8	2	2	4	3
1.78	3.25	1.78	4.69	9.42	8	2	3	3	4
1.42	8.00	11.25	13.93	1.42	24	9	8	7	5
1.96	5.30	1.96	3.16	11.82	4	1	0	3	6
2.61	5.11	2.61	4.00	12.08	4	0	0	4	7
0.47	7.35	11.02	13.86	0.47	24	8	8	8	8
1.10	8.49	12.21	15.03	1.10	26	9	8	9	9
3.25	3.25	6.87	9.70	7.90	14	2	3	9	10
1.25	1.25	4.30	7.35	7.56	12	2	5	5	11
1.79	7.72	11.16	13.93	1.79	24	8	9	7	12
0.47	7.35	11.02	13.86	0.47	24	8	8	8	13
2.00	6.71	2.48	2.00	12.99	2	0	2	0	14
1.03	1.03	3.63	6.71	7.98	11	2	4	5	15
3.80	6.71	10.02	12.69	3.80	21	8	4	9	16
1.68	8.98	12.75	15.59	1.68	27	9	9	9	17
1.25	1.25	4.30	7.35	7.56	12	2	5	5	18
3.29	6.33	3.29	4.00	12.24	4	0	4	0	19
0.79	8.19	11.70	14.46	0.79	25	9	8	8	20

Distance to the calculation in iteration 3 will do a comparison and have the shortest distance between the data center cluster, this range indicates that the data are in one group with the nearest cluster center. If there are no more clusters are not the same then the iteration process has been completed. In the data classification, 3rd iteration already represents the outcome of the clustering process carried out in the framework of profiling email content. The following will be displayed matrix data to the data classification 3rd iteration, the value of 1 means that the data resides in the data and a value of 0 means that are outside the data

Tabel 3.10 Grouping Data Iteration to 3

Cen 4	Cen 3	Cen 2	Cen 1	email
0	0	0	1	1
0	0	0	1	2
0	1	0	0	3
0	1	0	0	4
0	0	0	1	5
0	1	0	0	6
0	1	0	0	7
0	0	0	1	8
0	0	0	1	9
1	0	0	0	10
1	0	0	0	11
0	0	0	1	12
0	0	0	1	13
0	1	0	0	14
1	0	0	0	15
0	0	0	1	16
0	0	0	1	17
1	0	0	0	18
0	1	0	0	19
0	0	0	1	20

Table 3:10 above illustrates how a sample data taken for clustering process is carried out and the results obtained for each centroid. Of the 20 samples of the processed data generating centroid as follows:

a. Centroid one as much as 10 Data
b. Centroid 2 as much as 0 Data
c. Centroid 3 as much as 6 Data
d. Centroid 4 as much as 3 Data

4 RESULT

a. Clustering with RapidMiner

At this stage, the file used to cluster with RapidMiner is data_mailku.csv. This CSV file to be imported from RapidMiner then do clusters. The steps carried out in a cluster of email content with RapidMiner shown in Figure 6.

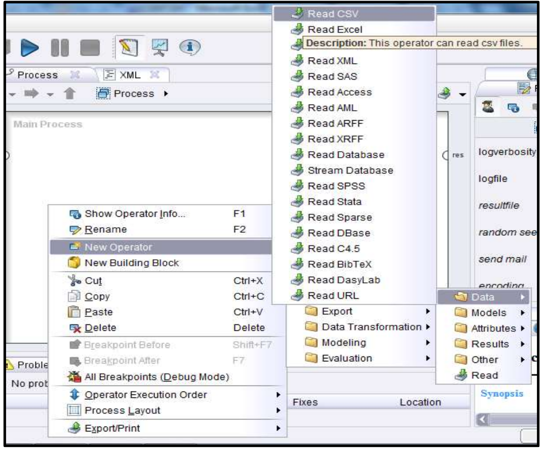


Figure 6 Importing file data-mailku.csv

After importing the file data_mailku.csv next steps for the attributes and cluster validation. On the Modeling tab window and Clustering and Segmentation Plug Read CSV Clustering operator as shown in Figure 7.

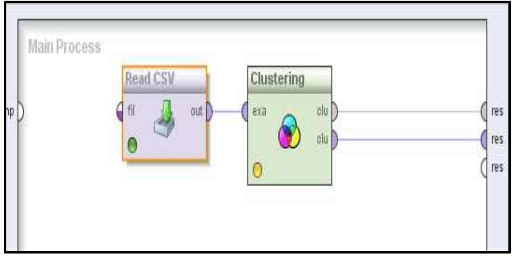
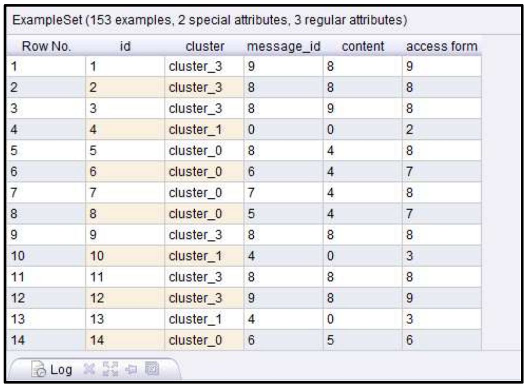


Figure 7. Connecting the CSV file with clustering

Set the number of clusters is required, the process is made 4 clusters consisting of True Mail, Promotion, Scam, and Unpurpose and click the Execute button. The results of the cluster can be seen in the Result Overview which can be displayed in a variety of perspectives. Figure 8 The following is the result of the cluster

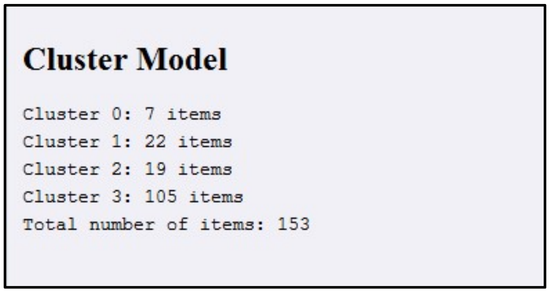


ExampleSet (153 examples, 2 special attributes, 3 regular attributes)

Row No.	id	cluster	message_id	content	access form
1	1	cluster_3	9	8	9
2	2	cluster_3	8	8	8
3	3	cluster_3	8	9	8
4	4	cluster_1	0	0	2
5	5	cluster_0	8	4	8
6	6	cluster_0	6	4	7
7	7	cluster_0	7	4	8
8	8	cluster_0	5	4	7
9	9	cluster_3	8	8	8
10	10	cluster_1	4	0	3
11	11	cluster_3	8	8	8
12	12	cluster_3	9	8	9
13	13	cluster_1	4	0	3
14	14	cluster_0	6	5	6

Figure 8: Result of the Cluster

Results in RapidMiner describe clustering cluster seen in Figure 9, the explanation for each cluster: Cluster 0 as Unpurpose, as Promotion Cluster 1, Cluster 2 as Fraud, and Cluster 3 as True email.



Cluster Model

```
Cluster 0: 7 items
Cluster 1: 22 items
Cluster 2: 19 items
Cluster 3: 105 items
Total number of items: 153
```

Figure 9. The clustering results by TextView

The clustering results can also be presented in various forms, following 10 images comparison of the forms of presentation of cluster process. One is a model that describes the pivot clustering email classification by type of cluster.

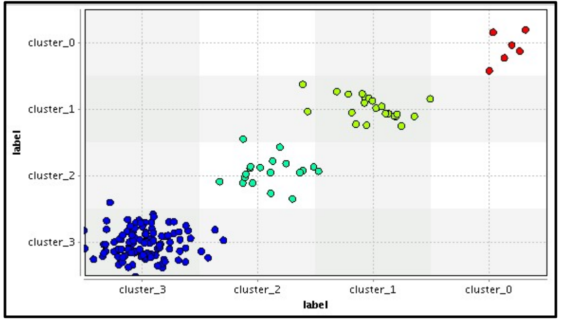


Figure 10: Results Clustering with Pivot view

b. Content profiling Email

From the results of email content above clustering, profiling will be conducted by the deductive method is to set the profile based on the data obtained from the scene, in this case, is the content of the email on the SMTP Mail Server. In the deductive method, the author using three main things:

1. Modus motive, in this case, analyzed how an

incoming email to the SMTP Mail Server and their characteristics. Referring to the clustering process that uses three criteria, namely:

a. Message ID, analysis is done by finding a common e-mail name format that is often used primarily Indonesian speaking, the more common email formats, it will have a bigger weight value
b. Access from, Analysis is done by finding the method used to access e-mail, for example, Imap, HTTP, telnet or the other. Access emails that have largest weight value are Imap because it is a commonly used method for accessing email.
c. Content, content analysis done by finding similarities string of email content itself compared to category generated email, including email true, scam/fraud, promotion, and unpurpose.
d. Signature, MO Although it can be different, but still there is one thing that must be present in every email access, in this case, the email address itself. Although email submissions via a relay that will disguise their email addresses, but still the main e-mail address will be detected. Signature is always associated with the perpetrator and is associated with motive. So, it will be unbelievably motive closely associated with Signature

2. Victimology
Analysis related to the victim, in this case, is a client of Postfix SMTP server, which in turn will impact on the performance of a machine that server. Of the 153-existing incoming mail on the server profiling results obtained include the category Unpurpose 4.58%, 14.38 Promotion categories, 12,42 Scam / Fraud categories and 68.63 True Email categories. For more details see the following table 4.11:

Tabel 4.11 Results in the percentage of email content classification

Percentage	Category	No
4.58%	UnPurpose	1
14.38%	Promotion	2
12.42%	Scam/Fraud	3
68.63%	True Email	4

Email the Postfix SMTP server cluster consisting of four illustrates that not all mail to the mail server desired by the client because there are some emails that are content just as trash/junk email that can interfere with the performance of the server.
Suggestions for further research is the amount of mail on the Postfix mail server is reproduced, it should be added elaboration of the email category unpurpose making kind would be more detailed and that the last should be made to the content of e-mail filtering interfere with the performance of Postfix SMTP Mail Server

5. CONCLUSION

The conclusion that has been gained from research on the analysis of email content with K-Means Clustering for process profiling Postfix SMTP server method successfully classify types of email content in accordance with the character, making can be used to filter the content of the unsolicited email, which primarily includes promotional category and fraud making the performance of Postfix SMTP Mail Server can work optimally. The stage profiling email content with K-Means Cluster Postfix SMTP server begins by capturing the log file from the SMTP server, and then analyzed K Clustering -Means using RapidMiner tool, and the next stage is profiling the content of the email. Profiling method used is the deductive method, namely setting up profiles based on data obtained from the scene, in this case, is the content of the email on the SMTP Mail Server. Algorithm K-Means Clustering can be used to analyze the types of email content on postfix mail server in order to obtain four categories of email, namely: True email, Scam / Fraud, Promotion and Unpurpuse by classifying each pattern characteristic of mail based on three criteria, namely message_id, Access from, and Content. 153 Data analyzed by K-Means Clustering and deductive profiling yield 4.58% are Unpurpose, 14.38 Promotion category, category 12:42 Scam / Fraud and 68.63 True Email category.

REFERENCES:

[1] A. Kurniawan, I. Riadi, and A. Luthfi, "Forensic Analysis and Prevent of Cross Site Scripting in Single Victim Attack Using Open Web Application Security Project (Owasp) Framework," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 6, pp. 1363–1371, 2017.

[2] E. Husni and A. Wibowo, "E-mail System for Delay Tolerant Network," 2012.

[3] S. Nizamani, N. Memon, M. Glasdam, and D. D. Nguyen, "Detection of fraudulent emails by employing advanced feature abundance," *Egypt. Informatics J.*, vol. 15, no. 3, pp. 169–174, 2014.

[4] F. A. Hermawati, *Data Mining*. Yogyakarta: Andi Offset, 2013.

[5] C. Amanpreet, M. Gaurav, and G. Kumar, "Survey on Data Mining Techniques in Intrusion Detection," *Int. J. Sci. Eng. Res.*, vol. 2, no. 7, pp. 2–5, 2011.

[6] T. Balasubramanian and R. Umarani, "An analysis on the impact of fluoride in human health (dental) using clustering data mining technique," in *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)*, 2012, pp. 370–375.

[7] A. Pathak, S. A. R. Jafri, and Y. C. Hu, "The case for spam-aware high performance mail server architecture," *Proc. - Int. Conf. Distrib. Comput. Syst.*, pp. 155–164, 2009.

[8] C. C. Aggarwal and C. K. Reddy, *DATA CLUSTERING Algorithms and Applications*. 2013.

[9] M. Bramer, "Introduction to Data Mining," 2016, pp. 1–8.

[10] P.-N. Tan, *Introduction to DATA MINING Second Edition*. Boston : Pearson Addison Wesley, 2013.

[11] I. Riadi, J. Eko, A. Ashari, and Subanar, "Log Analysis Techniques using Clustering in Network Forensics," *Int. J. Comput. Sci. Inf. Secur.*, vol. Vol. 10, N, no. August 2016, 2013.

[12] B. Ratner and D. Ph, *Statistical and Machine-Learning Data Mining*, no. 1908. CRC Press, 2013.

[13] W. Wu and M. Peng, "A Data Mining Approach Combining K-Means Clustering with Bagging Neural Network for Short-term Wind Power Forecasting," *IEEE Internet Things J.*, vol. 4662, no. c, pp. 1–1, 2017.

[14] K. Adi, "Detection of the Beef Quality," pp. 253–259, 2016.

[15] A. Al-Wakeel and J. Wu, "K-means based cluster analysis of residential smart meter measurements," *Energy Procedia*, vol. 88, pp. 754–760, 2016.

[16] D. Kulshreshtha, V. P. Singh, A. Shrivastava, A. Chaudhary, and R. Srivastava, "Content-Based Mammogram Retrieval Using k-means Clustering and Local Binary Pattern," pp. 634–638, 2017.

[17] N. K. M. Madi, S. Salehian, F. Masoumiyan, and A. Abdullah, "Implementation of secure email server in cloud environment," *2012 Int. Conf. Comput. Commun. Eng. ICCCE 2012*, no. July, pp. 28–32, 2012.

[18] O'Reilly, *Postfix: The Definitive Guide*. 2003.