E-ISSN: 2528 - 6544
P-ISSN: 2620 - 3383
Vol.4 No.2 February 2020

TMJ

Technomedia Journal

iLearning Journal Center (iJC)

AMFC : A Novel Archive Modeling Based On Data Cluster and Filtering

Rolly Maulana Awangga[1]
Syafrial Fachri Pane[2]
Cahya Kurniawan[3]

[1,2,3] Applied Bachelor Program of Informatics Engineering, Politeknik Pos Indonesia, Jl. Sariasih No.54, Sarijadi, Sukasari, Bandung 40151, Telp. (022) 2009562, Indonesia
Email: *awangga@poltekpos.ac.id*[1], *syafrial.fachri@poltekpos.ac.id*[2], *cahyakurniawan99@gmail.com*[3]

ABSTRAK

Pengarsipan file sekarang perlu dikelola dengan tepat sehingga mudah ditemukan dan dikelola. Pengarsipan file yang dimaksud adalah bagaimana membantu dalam proses pencarian data dengan jumlah yang cukup banyak, untuk memudahkan pekerjaan mengikuti tujuan untuk mengurangi waktu pencarian data dapat diintegrasikan dengan sistem yang dibuat. Pengarsipan itu sendiri bertujuan untuk memudahkan pengelolaan data yang sangat beragam dan dengan jumlah yang besar, untuk memudahkan pengelolaan dan juga pengendalian yang dilakukan. Masalah dengan pengarsipan arsip, dalam hal ini, adalah kurangnya manajemen mengenai pengarsipan arsip yang benar. Manajemen file arsip yang dikontrol membuat jumlah waktu yang dilewati hanya dengan mencari file. Dengan melihat sejumlah besar data, perlu untuk menggunakan metode untuk dapat mengelola dan mencari, dengan Alphabet, Numerik dan metode K-means Clustering dirancang dan diproses, lebih mudah untuk mengelola pencarian data yang ada sehingga pekerjaan tidak t terlalu banyak waktu. Perlu ada pengembangan lebih lanjut dari analisis yang dilakukan saat ini untuk lebih meningkatkan efektivitas dengan menciptakan sistem mengikuti aturan analisis yang dibuat saat ini.

Kata Kunci: *Pengarsipan, Arsip, Manajemen File Sistem Pengisian Abjad, Sistem Pengarsipan Numerik, K-means Clustering.*

ABSTRACT

File archiving now needs to be appropriately managed so that it is easy to find and manage. File archiving in question is how to help in the process of finding data with a considerable number, to facilitate the work following the aim to reduce the time the search data can be integrated with the system created. Archiving itself aims to facilitate the management of data that is very diverse and with a large amount, to facilitate the management and also the control carried out. The problem with filing archives, in this case, is the lack of management regarding the correct filing of archives. Controlled archive file management makes the amount of time that is passed only by searching files. By looking at large amounts of data, it is necessary to use methods to be able to manage and search, with Alphabet, Numeric and K-means Clustering methods designed and processed, it is easier to manage *existing data searches so that the work doesn't take too much time. There needs to be further development of the analysis carried out at this time to further improve the effectiveness by creating a system following the analysis rules made at this time.*

Keywords: *Filing, Archives, File Management Alphabetical Filling System, Numerical Filing System, K-means Clustering.*

INTRODUCTION

Information is a topic of discussion identified in a media or file (Lyman et al., 2016). Where technology is now built with the aim of reducing a complicated job because the information plays a key (Gandomi & Haider, 2015) The application of information technology to companies, offices, and organizations are viewed as one of the solutions that will be able to increase the competence level (Sihotang, 2015).

ATR / BPN Sidoarjo is a non-ministerial government institution in Indonesia are obliged to carry out tasks in the area of land administration in accordance with the provisions of the law are isolated. Moreover, have many files on the formulation of policies on land use, drafting, and implementation of systems in the field of surveying, measuring and mapping. So with so many files on this data, we need good management to make. So the problem with offices ATR / BPN Sidoarjo is an issue which is the process of archiving the archive for data that is still done manually and some archives are yet sorted into data sought or not. Many think that managing files is very heavy, but all of it is just the opposite of that statement, and its management is represented by objects and data(Cavus & Zabadi, 2014). With this archive, management can be useful in the future (Oltmans, van Diessan, & van Wijngaarden, 2014). Records management is done to cut solution to maintain diversity among solutions in the archives (Nag, Pal, & Pal, 2015) By managing the archiving of records, it is easy for workers to carry out activities related to managing their files. the actual physical data is very important. Archives must also be maintained in the information using a coding standard that is easily understood and is also driven by the requirements of images owned (Clark et al., 2014).

In the archive management records, it is necessary to develop and improve the performance of the work(Pane, Awangga, & Azhari, 2018), in which the study data clustering is required(R. M. Awangga, Pane, Tunnisa, & Suwardi, 2018). Analysis of the use of statistical analysis methods alphabet (R. Awangga, 2017). Alphabet and numerical methods applied to aim to initialization coded data and also the K-means clustering which serves as a grouping with a specified number of clusters (Jumb, Sohani, & Shrivas, 2014). Management of archive files provides data support for the functions of a search system (Zhang, Yang, & Song, 2015) The code in the alphabetical method identifies the data specified so that the management did not pose a problem (Odeh, Abu-Errub, & Awad, 2015). Alphabetical and Numerical Methods Filling System method used to form the unique character set of the data (Nigam & Sahu, 2015). In the K-means clustering useful to split the data into several groups to determine the cluster (Rajalakshmi, Dhenakaran, & Roobin, 2015).

Analysis of these activities are activities such as parsing, separating, sorting data or some things that can be used to create a new archive. In the use of methods which do aims to make it more efficient in its management (Baladhandabany, Gowtham, Kowsikkumar, Gomathi, & Vijayasalini, 2015). Archiving data requires data that includes data distribution initialization data for faster and more accurate (Mashilkar, Khaire, & Dalvi, 2015).

In this analysis the necessary selection and consideration of several methods in accordance with what was discussed. The first consideration regarding serial dilute method where this method focuses on repetition completely randomized design. And also multiclass classification method that focuses on the observation and the observation data analyzed new data (Lalis, 2016). Methods RBAC (Role Based Access Control) where this method of management authority (Lin, Li, & Ma, 2014). Several methods mentioned above having a base and a different principle (Jayalakshmi & Pandian, 2014). Using the methods applied in this analysis will illustrate the efficiency and accuracy appropriate to the data (Borhanifar & Sadri, 2014).

Management information is used to provide information services (Kostoska, Chorbev, & Gusev, 2014). In their records management needs of collaboration and coordination(Poba-Nzaou, 2016). In this archive filing arrangement necessary requirements and conditions set forth in order to be well controlled (Klampfl, Granitzer, Jack, & Kern, 2014). So that becomes centralized management (Gospodinov, Gospodinova, & Cheshmedijev, 2014). Differences of several international journals are bound or be references regarding the effectiveness of its use. Where the use of methods of the alphabet and the numeric system is used to initialization and manufacture new code (Nyers, Garbai, & Nyers, 2014) . And also the method of K-means clustering function to a dataset into clusters(Cebeci & Yildiz, 2015). Where the system settings to a file named and logically placed to be able to save easily. This method also controls how information is stored and retrieved.

The literature review sought is used to understand the topic to be discussed. Journal with the title "Web-Based Document Filing Information System". Explain the records in the company's business processes that use the PHP programming language with the results of the Academic and Student Administration Bureau archive report differences with the research conducted is the use of programming languages where the research actually uses PHP while researchers use Codeigniter Language. Then the journal with the title "Information Systems On Archives In State Gembong Kab. Multi User-Based Pati ". In this journal discusses archiving, which still uses manual methods for the main archiving media, so that this journal produces an information system about archiving. the difference made by this research is how the management of the archive itself, where the journal manages an information archive system that is only made according to the flow of desires, but in this study the application of the K-Means Clustering method is applied for its management so that it is easier to group the types archive in it.

THE PROBLEM

In the problem that was done in this study was the absence of storage and management of data / archives at the ATR / BPN Sidoarjo office, so management was needed, and here researchers used alphabetical and numerical methods and kmeans clustering. then the method is applied in making the application where later the results will be expected to schedule the problems that exist in the ATR /

BPN Siodarjo office.

RESEARCH METHOD

1. Troubleshooting

   On the search for problems that are done first by preparing what will be asked later. where the problem search with interviews is a sam- pling technique that will be analyzed and tested and the way that is done with interviews is to ask directly to the external counselor about matters relating to what will be tested and what needed. After conducting a problem search the next step is data collection.

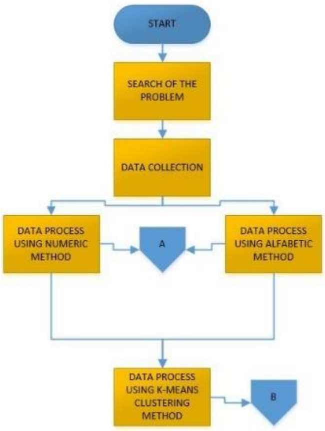

START
SEARCH OF THE PROBLEM
DATA COLLECTION
DATA PROCESS USING NUMERIC METHOD
A
DATA PROCESS USING ALFABETIC METHOD
DATA PROCESS USING K-MEANS CLUSTERING METHOD
B

Figure 3.1: Process Data

2. Data Collection

   In the data collection in question is how to get the data to be managed, and how to request raw data directly to the external counselor where later the raw data will be used to manage the data . when you have obtained the next step is needed to process the data using the alphabetical method.

3. Process Data Alphabetically and Numerically

   a. Raw Data

      In the process of managing data there needs to be data to be managed, where the raw data produced is used for the process of applying the method. And the raw data in question is data generated from data collection results.

   b. Application Of Methods

      The process of applying the method is an advanced process of the results of previous raw data management which is useful for reference management to be carried out. And the process of applying the method to be used is in accordance with the rules and agreements that have been previously confirmed. And then the process is done using alphabetical and numeric methods

   c. Alphabetical Method

The use of the alphabetical method is used to find alphabetical data wherein the purpose is to obtain existing alphabetical data. And the method used refers to the previous process, namely collecting data. So that raw data can be managed in this method. And the equation used in this method can be seen in equation 1 with an explanation of the equation as below.

$$F(x) = (E2\{A2\}\{B8\})\ 2 \wedge 0$$

The equation above is used to manufacture a new code based on the reference on

village data with the following explanation:

In the process the existing raw data is managed by using equation 1 in validation whether the process is correct or not, if not then the process is done again when it is correct then the alphabetical data is obtained and the subsequent process will be carried out using the K-Means Clustering method.

d. Numerical Method

The use of this numerical method is used to search for alphabetical data where the purpose is to obtain existing numerical data. And the method used refers to the previous process, namely collecting data. So that raw data can be managed in this method. And the equation used in this method can be seen in equation 2 with an explanation of the equation as below:

$$F(x) = (G2\{C\ 2\}\{D5\})\ 2 \wedge 0$$

The equation above is used to manufacture a new code based on the reference on

village data with the following explanation:

In the process of the existing raw data managed by using equationin validation whether the process is correct or not, if not then the process is done again if it is correct then numerical data has been obtained then the process will be carried out using the K-Means Clustering method.

e. Application of K-Means Clustering Method

The process of the implementation of the method of K-Means Clustering is the process of grouping data by using a equation on the K-Means Clustering which will be described in more detail in K-Means Clustering Method.

4. K-Means Clustering Method

This method is used to separate in agglomeration data sets into smaller groups. This method is used by selecting some of the data of all the data.Where the group/groups are formed by minimizing the amount of data of the Euclidean distance between the data and its center point.

a. Alphabetical and Numeric Data

alphabetical and numerical data obtained from the process of using alphabetical and numerical methods will be processed using the application of the K-Means Clustering method where data management is taken from managed data using alphabetical methods and numerical methods, where data will be processed using stages according to K-means Clustering rules. the first step that is done is initializing the data and can be seen in the data initialization process.

b. initialization Data

The process of initiating this data uses the excel equation which is the vlookup equation, which previously used the equation to make references from the alphabet and numeric methods

$$\mathrm{F}\,(\mathrm{x}) = (\mathrm{E9}\,\{\mathrm{C}\,2\}\,\{\mathrm{D5}\})\;2 \wedge 0$$

The use of the equation above in Table districts, with the following caption:

a. E9 = Data sub-district
b. C2 = Reference the data sub-district
c. D5 = Code sub-district

$$\mathrm{F}\,(\mathrm{x}) = (\mathrm{D9}\,\{\mathrm{E2}\}\,\{\mathrm{F}\,8\})\;2 \wedge 0$$

The use of the equation above the village table, with the following caption:

a. D9 = Data villages
b. E2 = Reference of villages
c. F8 = Village code

c. Determination of Number of Clusters

This process determines the number of clusters that will be used by the random selection, random selection are already established in the rules of the use of K-Means method.

d. Calculations Distance Object To Cluster
e. Grouping Based on Shortest Distance
f. Rate Centroid by Random
g. Update latest Value
h. Checking Value Cluster
i. Iteration
j. Results Data

RESULTS AND DISCUSSION

1. Numerical and Alphabetical Methods

   The use of this method to present data efficiently to find a significant parameter (Haque, Amale, & Kamble, 2014). The alphabetic method adopted in the dictionary of terms or be referred by initialization the data (Abbas & Yasin, 2016). Numerical methods are used as initialization numbers with the specified parameters. In using the above method in which the initialization need to establish appropriate data sorting (Kiss, Genge, Haller, & Sebestyén, 2014).

2. K-Means Clustering Data

This method is used to separate in agglomeration data sets into smaller groups. This method is used by selecting some of the data of all the data (Wang et al., 2015). Where the group/groups are formed by minimizing the amount of data of the Euclidean distance between the data and its center point.
workflow as follows:

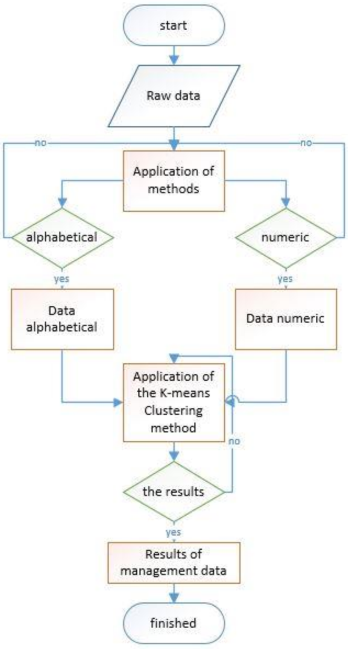


Figure 1 data management process

The raw data will be managed using the alphabet and numerical methods followed by a K-means cluster management with the aim to eliminate redundancies and reduce a large amount of data (Harb, Makhoul, & Couturier, 2015). Next will be calculated and where the existing grouping of data disaggregated clusters (Puzyn, Mostrag-Szlichtyng, Gajewicz, & Skrzyński Michałand Worth, 2014). And also randomly determined cluster centers (Li, Song, Wei, Lu, & Zhu, 2015).

The parameters to be entered when using K-means cluster algorithm is a village and district in which the steps start from using equation in the below :

$$\mathrm{F(x)} \;=\; (\mathrm{lv}\,\{\mathrm{ta}\}\,\{\mathrm{colindex}\})\ \mathrm{rl}\ \wedge\ 0$$

$$\left[\frac{amount\ of\ data}{number\ of\ clusters+1}\right]$$

$$\sqrt{(d^2-i^3)^2+(e^4-j^3)^2}$$

a. The use of the formula 1 is for the numeric and alphabetic method in which to excel formula will produce initialization each set of data based on the code, followed by the use of the second formula wherein this formula 2 is used to select k data as a centroid. After that calculate the shortest distance to each cluster using a formula 3 formula d2 into sub-districts code has been initialized, i3 which is a random cluster on c1, e4 which the village code has been initialized and j3 which is a random cluster on c1. Then to find the value of c2 keep using the formula 3 only values i3 and i4 replaced j3 and j4 which is a random cluster on c2, and so on c3 just change the value of the cluster on c3.

$$\Rightarrow \left(\frac{m4}{o4}\right) \Rightarrow m4 \cap \Rightarrow \left(\frac{m4}{04}\right) \Rightarrow n4 \ \cap \Rightarrow \left(\frac{m4}{04}\right) = 04$$

b. Then the formula 4 is used as the value of the membership of each cluster, where the value is the value of cluster c1 m4, o4 value of cluster c3, m2 value c1, c2 cluster value n4, n2 value c2, c3 o2 value. Then after getting the value of membership in the sought value of the minimum distance using formula 5. And also the minimum distance squared value using the formula 6.

$$s = \frac{m4}{04}$$

$$q4 * q4$$

c. In Formula 7 It Is Used To Find The Value Of Cluster 1 In The Sub-District Column. Where If In The Membership Column (P4) Is Equal To C1 Then Fill With The Value That Is In The Sub-District Code Column (D2), And If Not Then There Is No Result. And In Formula 8 The Explanation Is The Same As The Previous Formula Where The Change Only In Filling In The Value Is Changed To The District Code (E2).

$$\Rightarrow p4 = c1 \Rightarrow d2$$

$$\Rightarrow p4 \neq c1 \Rightarrow e2$$

d. In Formula 9 It Is Used To Find The Value Of Cluster 1 In The Sub-District Column. Where If In The Membership Column (P4) Is Equal To C2, Then Fill In The Value That Is In The Sub-District Code Column (D2), And If Not Then There Is No Result. And In Formula 10 The Explanation Is The Same As The Previous Formula Where The Change Only In The Filling Of The Value Is Changed To The District Code (E2).

$$\Rightarrow p4 = c2 \Rightarrow d2$$

$$\Rightarrow p4 \neq c2 \Rightarrow e2$$

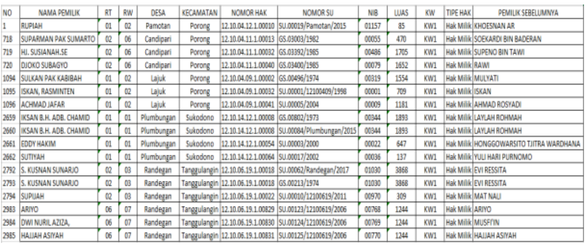
e. In formula 11 it is used to find the value of cluster 1 in the sub-district column. Where if in the membership column (p4) is equal to c3 then fill in the value that is in the sub-district code column (d2), and if not then there is no result. And in formula 12 the explanation is the same as the previous formula where the change only in filling in the value is changed to the district code (e2).

$$\Rightarrow p4 = c3 \Rightarrow d2$$

$$\Rightarrow p4 \neq c3 \Rightarrow e2$$

3. Data Source

   In this study, data sources took on PTSL project data in ATR / BPN Sidoarjo office and what is shown only part of the available data. and can shown in Figure 2



NO	NAMA PEMILIK	RT	RW	DESA	KECAMATAN	NOMOR HAK	NOMOR SU	NIB	LUAS	KW	TIPE HAK	PEMILIK SEBELUMNYA
1	RUPIAH	01	02	Pamotan	Porong	12.10.04.12.1.00010	SU.00019/Pamotan/2015	01157	85	KW1	Hak Milik	KHOESNAN AR
718	SUPARMAN PAK SUMARTO	02	06	Candipari	Porong	12.10.04.11.1.00013	GS.03003/1982	00055	470	KW1	Hak Milik	SOEKARDI BIN BADERAN
719	HJ. SUSIANAH.SE	02	06	Candipari	Porong	12.10.04.11.1.00032	GS.03392/1985	00486	1705	KW1	Hak Milik	SUPENO BIN TAWI
720	DJOKO SUBAGYO	02	06	Candipari	Porong	12.10.04.11.1.00040	GS.03400/1985	00079	1652	KW1	Hak Milik	RAWI
1094	SULKAN PAK KABIBAH	01	02	Lajuk	Porong	12.10.04.09.1.00002	GS.00496/1974	00319	1554	KW1	Hak Milik	MULYATI
1095	ISKAN, RASMINTEN	01	02	Lajuk	Porong	12.10.04.09.1.00032	SU.00001/12100409/1998	00001	709	KW1	Hak Milik	ISKAN
1096	ACHMAD JAFAR	01	02	Lajuk	Porong	12.10.04.09.1.00041	SU.00005/2004	00009	1181	KW1	Hak Milik	AHMAD ROSYADI
2659	IKSAN B.H. ADB. CHAMID	01	01	Plumbungan	Sukodono	12.10.14.12.1.00008	GS.00802/1973	00344	1893	KW1	Hak Milik	LAYLAH ROHMAH
2660	IKSAN B.H. ADB. CHAMID	01	01	Plumbungan	Sukodono	12.10.14.12.1.00008	SU.00084/Plumbungan/2015	00344	1893	KW1	Hak Milik	LAYLAH ROHMAH
2661	EDDY HAKIM	01	01	Plumbungan	Sukodono	12.10.14.12.1.00054	SU.00003/2000	00022	647	KW1	Hak Milik	HONGGOWARSITO TJITRA WARDHANA
2662	SUTIYAH	01	01	Plumbungan	Sukodono	12.10.14.12.1.00064	SU.00017/2002	00036	137	KW1	Hak Milik	YULI HARI PURNOMO
2792	S. KUSNAN SUNARJO	02	03	Randegan	Tanggulangin	12.10.06.19.1.00018	SU.00062/Randegan/2017	01030	3868	KW1	Hak Milik	EVI RESSITA
2793	S. KUSNAN SUNARJO	02	03	Randegan	Tanggulangin	12.10.06.19.1.00018	GS.00213/1974	01030	3868	KW1	Hak Milik	EVI RESSITA
2794	SUPIJAH	02	03	Randegan	Tanggulangin	12.10.06.19.1.00022	SU.00010/12100619/2011	00970	309	KW1	Hak Milik	MAT NAJI
2983	ARIYO	06	07	Randegan	Tanggulangin	12.10.06.19.1.00829	SU.00123/12100619/2006	00768	1244	KW1	Hak Milik	ARIYO
2984	DWI NURIL AZIZA,	06	07	Randegan	Tanggulangin	12.10.06.19.1.00830	SU.00124/12100619/2006	00769	1244	KW1	Hak Milik	MUSFI'IN
2985	HAJJAH ASIYAH	06	07	Randegan	Tanggulangin	12.10.06.19.1.00831	SU.00125/12100619/2006	00770	1244	KW1	Hak Milik	HAJJAH ASIYAH

Figure 2 raw data

4. Data Processing

   Before the data are grouped according to the criteria, the raw data is converted by initializing the data into a number using the alphabet and numerical methods. And also from the raw data, only 2 data is taken as the testing center, the villages, and districts. Figure 1 is the result of initialization using the alphabet and numerical methods by using formula 1.

Table 1 Initialization results use alphabetical and numeric methods

No.	village	sub-district	sub-district code	village code
1	Pamotan	Porong	1	1
718	Candipari	Porong	1	1
719	Candipari	Porong	1	1
720	Candipari	Porong	1	1
1094	Lajuk	Porong	1	1
1095	Lajuk	Porong	1	1
1096	Lajuk	Porong	1	1
2659	Plumbungan	Sukodono	2	4
2660	Plumbungan	Sukodono	2	4
2661	Plumbungan	Sukodono	2	4
2662	Plumbungan	Sukodono	2	4
2792	Kebakalan	Porong	1	1
2793	Kebakalan	Porong	1	1
2794	Kebakalan	Porong	1	1
2983	Jeruklegi	Balongbendo	4	6
2984	Jeruklegi	Balongbendo	4	6
2985	Jeruklegi	Balongbendo	4	6

Table 2 cluster determination

	Cluster Center	Random	
C1	2	1	1
C2	2659	2	4

C3	2983	4	6

In table 1 is the result of initializing data using alphabetical and numerical methods. Based on the ranges specified in the data ptsl such as table 2, it can be concluded that the grouping of data ptsl to the assessment center without using the method but random selection.

5. Process data

Once the data is processed, the next step is data processed to form a group of data. Modified data will be processed using the k- means clustering. And steps in accordance with figure 3.

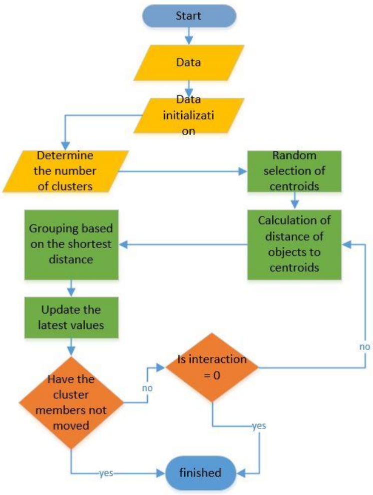

Start
Data
Data initialization
Determine the number of clusters
Random selection of centroids
Calculation of distance of objects to centroids
Grouping based on the shortest distance
Update the latest values
Is interaction = 0
no
Have the cluster members not moved
no
yes
yes
finished

Figure 3 process data

At this stage, the raw data will be processed initialized using alphabetic and numeric methods in Excel using formulas vlookup like in formula 1, then determine the number of clusters selected at random. Then do the calculation of each object to the centroid distance using the formula in section method, and the following data is generated as in Table 3.

Table 3 the center distance to the cluster

	the center distance to the cluster		
cluster center	C1	C2	C3
	2	2659	2983
1	0	3,162278	5,830952
718	0	3,162278	5,830952
719	0	3,162278	5,830952
720	0	3,162278	5,830952
1094	0	3,162278	5,830952
1095	0	3,162278	5,830952
1096	0	3,162278	5,830952

2659	3,162278	0	2,828427
2660	3,162278	0	2,828427
2661	3,162278	0	2,828427
2662	3,162278	0	2,828427
2983	5,830952	2,828427	0
2984	5,830952	2,828427	0
2985	5,830952	2,828427	0

Then the placement of the membership is done, at least the minimum squared distance and from a distance so that any data in any cluster. With the formula in the formula 4, formula 5, formula 6. Thus generating the data in Table 4.

Table 4 membership placement, minimum distance and minimum square of distance

the center distance to the cluster						
cluster center	C1	C2	C3	membership	min distance	min square distance
	2	2659	2983			
1	0	3,162278	5,830952	C1	0	0
718	0	3,162278	5,830952	C1	0	0
719	0	3,162278	5,830952	C1	0	0
720	0	3,162278	5,830952	C1	0	0
1094	0	3,162278	5,830952	C1	0	0
1095	0	3,162278	5,830952	C1	0	0
1096	0	3,162278	5,830952	C1	0	0
2659	3,162278	0	2,828427	C2	0	0
2660	3,162278	0	2,828427	C2	0	0
2661	3,162278	0	2,828427	C2	0	0
2662	3,162278	0	2,828427	C2	0	0
2983	5,830952	2,828427	0	C3	0	0
2984	5,830952	2,828427	0	C3	0	0
2985	5,830952	2,828427	0	C3	0	0

Then the new cluster center grouping is done using the formula 7, formula 8, formula 9, formula 10, formula 11, formula 12 and produce the data in Table 5.

Table 5 new cluster center

C1		C2		C3	
sub-district	village	sub-district	village	sub-district	village
1	1				
1	1				
1	1				
1	1				

1	1				
1	1				
1	1				
		1	1		
		1	1		
		1	1		
		1	1		
1	1				
1	1				
1	1				
				1	1
				1	1
				1	1

Then calculate the distance between the center of the cluster between c1 and c2, c1 to c3, and c2 to c3. The model is the same model but here formula 15, formula 16, formula 17,formula uses the value of each cluster. So it can be seen within the cluster center values in Table 6.

$$\sqrt{(1-2)^2+(1-4)^2}$$

In equation 13 it is explained about the cluster distance formula where the root value of cluster 1 is subtracted by the value of cluster 2 with the power of 2 so that the value in table 6 is generated.

$$(1-4)^2+(1-6)^2$$

In equation 14 it is explained about the cluster distance formula where the root value of cluster 1 is subtracted by the value of cluster 3 with the power of 2 so that the value in table \ref{tabel6} is generated.

$$\sqrt{(2-4)^2+(4-6)^2}$$

In equation 15 it is explained about the cluster distance formula where the root value of cluster 2 is subtracted by the value of cluster 3 with the power of 2 so that the value in table 6 is generated. And the sum of all distances between the cluster centers.

$$\sum(\boldsymbol{ac2}+\boldsymbol{ac3}+\boldsymbol{ac4})$$

Table 6 Distance between Cluster Centers

C1	C2	3,16227766
C1	C3	5,830951895
C2	C3	2,828427125
	BCV	11,82165668

Then determine the ratio. Ratio formula

$$\Rightarrow \left(\boldsymbol{r3119}\left(\frac{\boldsymbol{ac5}}{\boldsymbol{r3119}}\right)\wedge\boldsymbol{0}\right)$$

Here the ratio results equal to 0 so that the interaction does not need to be continued for grouping data can already be found.

And the results of the existing data in tables 1 - 7 above can be included in the application RapidMiner for testing and produced.

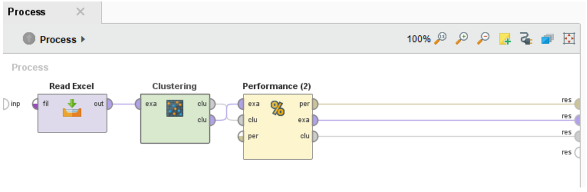


Figure 4 rapid miner process

6. Needs Analysis

The first phase in this analysis is the analysis of needs, which will be a key requirement needed for the process that will assist in processing. Here it will be reviewed whether the raw data will be compiled in accordance with the expected demand or not. It is made to work in a future position error does not occur.

7. Management data

The data in this analysis using three methods, the method of grouping by alphabetical, numerical and K-means. everywhere has its own steps so that the process becomes better management and support the expected demand.

8. Data processing using the K-Means Clustering method

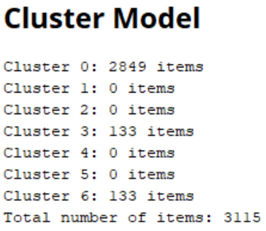
In the data K-Means Clustering algorithm that has been initialized using numerical and alphabetical method. The data will then be grouped in the process by specifying the desired cluster. Then calculate the distance in the center of the cluster group selected at random using the formula 3.

Then specify the gelatin find the placement of any data in any group using a formula that will seek the minimum distance and the squares of the distances. After performing results, the center of the cluster will be determined in order to generate value for the distance between the centers of clusters will produce the BCV which is used to find the value of the ratio. Where the value of this ratio to determine whether the data will be back or not. The results of the measurements in the table will be done correctly using rapid miner K-Means.

Table 7 The Result Data

Distance between cluster centers			Sub-district	Amount of data
C1	C2	3,16227766	Porong	2849
C1	C3	5,830951895	Sukodono	133
C2	C3	2,828427125	Balongbendo	133
		Bcv	11,82165668	
Ratio				0

And the results of RapidMiner can see in the table 8, the result is the same as the amount of data in accordance with the amount of previous data and can be seen from Figure 6.



Cluster Model

```
Cluster 0: 2849 items
Cluster 1: 0 items
Cluster 2: 0 items
Cluster 3: 133 items
Cluster 4: 0 items
Cluster 5: 0 items
Cluster 6: 133 items
Total number of items: 3115
```

Figure 5 data management process

CONCLUSION

In this experiment get the appropriate results to answer each identification problem where file archive data in the ATR / BPN Sidoarjo office can be managed appropri- ately and in accordance with desired expectations. File management using alphabet- ical, numerical and K-means grouping methods can help control data so that it can be seen in experiments.

because in its management it can be shown with real and true results.

In the application, it strongly supports the results expected to produce a clear grouping of data in accordance with the tests that have been carried out. so that after all repairs and management are carried out, the results can be used as new references for application in the form of applications and can also be applied to case studies on file management.

SUGGESTION

Suggestions for further research developers may be to manage operational vehicles and also grouping archives using Android, which can be easier and simpler to use and does not need to be managed using a computer or laptop media.

REFERENCE

[1] Abbas, A. A.-H., & Yasin, A. (2016). The Methodology of Creating Terminology in Rhetoric and Criticism Terms Dictionaries. *US-China Foreign Language*, 79.

[2] Awangga, R. (2017). Pengajuan Model Pengambilan Data pada Sistem Pemilu di Indonesia. *Jurnal Teknik Informatika*, *9*(1), 1–7. Retrieved from https://jurnal.diplomainformatika.or.id/teknikinformatika/article/view/9

[3] Awangga, R. M., Pane, S. F., Tunnisa, K., & Suwardi, I. S. (2018). K Means Clustering and Meanshift Analysis for Grouping the Data of Coal Term in Puslitbang tekMIRA. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *16*(3).

[4] Baladhandabany, D., Gowtham, S., Kowsikkumar, T., Gomathi, P., & Vijayasalini, P. (2015). PLC based automatic liquid filling system. *International Journal of Computer Science and Mobile Computing*, *4*(3), 684–692.

[5] Borhanifar, A., & Sadri, K. (2014). Numerical solution for systems of two dimensional integral equations by using Jacobi operational collocation method. *Sohag J. Math*, *1*(1), 15–26.

[6] Cavus, N., & Zabadi, T. (2014). A comparison of open source learning management systems. *Procedia-Social and Behavioral Sciences*, *143*, 521–526.

[7] Cebeci, Z., & Yildiz, F. (2015). Comparison of K-means and Fuzzy C-means algorithms on different cluster structures. *AGRÁRINFORMATIKA/JOURNAL OF AGRICULTURAL INFORMATICS*, *6*(3), 13–23.

[8] Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., … others. (2014). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of Digital Imaging*, *26*(6), 1045–1057.

[9] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

[10] Gospodinov, M., Gospodinova, E., & Cheshmedijev, K. (2014). SCADA System for Management and Visualization in Generation of Renewable Energy from Biomass. *International Journal of Engineering Research & Technology*, *3*(5), 882–887.

[11] Haque, A., Amale, A. B., & Kamble, P. D. (2014). Multiple Optimization of Wire EDM Machining Parameters Using Grey Based Taguchi Method for Material HCHCR. *Int. J. Mod. Eng. Res*, *4*(2), 39–45.

[12] Harb, H., Makhoul, A., & Couturier, R. (2015). An enhanced K-means and ANOVA-based

clustering approach for similarity aggregation in underwater wireless sensor networks. *IEEE Sensors Journal*, *15*(10), 5483–5493.

[13] Jayalakshmi, M., & Pandian, P. (2014). A method for solving quadratic programming problems having linearly factorized objective function. *International Journal Of Modern Engineering Research*, *4*, 20–24.

[14] Jumb, V., Sohani, M., & Shrivas, A. (2014). Color image segmentation using K-means clustering and Otsus adaptive thresholding. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *3*(9), 72–76.

[15] Kiss, I., Genge, B., Haller, P., & Sebestyén, G. (2014). Data clustering-based anomaly detection in industrial control systems. In *Intelligent Computer Communication and Processing (ICCP), 2014 IEEE International Conference on* (pp. 275–281).

[16] Klampfl, S., Granitzer, M., Jack, K., & Kern, R. (2014). Unsupervised document structure analysis of digital scientific articles. *International Journal on Digital Libraries*, *14*(3–4), 83–99.

[17] Kostoska, M., Chorbev, I., & Gusev, M. (2014). Creating portable TOSCA archive for iKnow university management system. In *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on* (pp. 761–768).

[18] Lalis, J. T. (2016). A New Multiclass Classification Method for Objects with Geometric Attributes Using Simple Linear Regression. *IAENG International Journal of Computer Science*, *43*(2), 198–203.

[19] Li, X., Song, K., Wei, G., Lu, R., & Zhu, C. (2015). A novel grouping method for lithium iron phosphate batteries based on a fractional joint Kalman filter and a new modified K-means clustering algorithm. *Energies*, *8*(8), 7703–7728.

[20] Lin, L.-L., Li, L.-Y., & Ma, Y. (2014). Design of data archive in virtual test architecture. *Journal of Information Hiding and Multimedia Signal Processing*, *5*(1), 80–89.

[21] Lyman, G. H., Somerfield, M. R., Bosserman, L. D., Perkins, C. L., Weaver, D. L., & Giuliano, A. E. (2016). Sentinel lymph node biopsy for patients with early-stage breast cancer: American Society of Clinical Oncology clinical practice guideline update. *Journal of Clinical Oncology*.

[22] Mashilkar, B., Khaire, P., & Dalvi, G. (2015). Automated bottle filling system. *International Research Journal of Engineering and Technology (IRJET)*, 56–2395.

[23] Nag, K., Pal, T., & Pal, N. R. (2015). ASMiGA: An Archive-Based Steady-State Micro Genetic Algorithm. *IEEE Transactions on Cybernetics*, *45*(1), 40–52. https://doi.org/10.1109/TCYB.2014.2317693

[24] Nigam, J., & Sahu, S. (2015). Fast and Effective System for Identifying Entity Names in Big Data. *Int. Journal of Computer Science and Engineering*, *3*.

[25] Nyers, J., Garbai, L., & Nyers, A. (2014). Analysis of heat pump condenser’s performance using the mathematical model and a numerical method. *Acta Polytechnica Hungarica*, *11*(3).

[26] Odeh, A., Abu-Errub, A., & Awad, M. (2015). Symmetric Key Generation Method using Digital Image. *International Journal of Computer Science Issues (IJCSI)*, *12*(2), 254.

[27] Oltmans, E., van Diessan, R. J., & van Wijngaarden, H. (2014). Preservation functionality in a digital archive. In *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004.* (pp. 279–286). https://doi.org/10.1109/JCDL.2014.240049

[28] Pane, S. F., Awangga, R. M., & Azhari, B. R. (2018). Qualitative Evaluation of RFID Implementationon Warehouse Management System. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *16*(3).

[29] Poba-Nzaou, P. (2016). Electronic Health Record in Hospitals: A Theoretical Framework for Collaborative Lifecycle Risk Management. *Journal of Healthcare Communication*, *1*, 2.

[30] Puzyn, T., Mostrag-Szlichtyng, A., Gajewicz, A., & Skrzyński Michałand Worth, A. P. (2014). Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. *Structural Chemistry*, *22*(4), 795–804.

[31] Rajalakshmi, K., Dhenakaran, D. S. S., & Roobin, N. (2015). Comparative Analysis of K-Means Algorithm in Disease Prediction. *International Journal of Science, Engineering and Technology Research (IJSETR)*, *4*(7), 1–3.

[32] Sihotang, H. T. (2015). Implementation of Information Technology Governance Using Cobit Framework 4.1 Case Study at PT. Perkebunan Nusantara III Medan (Persero). *Journal of Mantik Penusa*.

[33] Wang, J., Wang, J., Song, J., Xu, X.-S., Shen, H. T., & Li, S. (2015). Optimized cartesian k-means. *IEEE Transactions on Knowledge & Data Engineering*, (1), 1.

[34] Zhang, W., Yang, X., & Song, Q. (2015). CONSTRUCTION OF TRACEABILITY SYSTEM FOR MAINTENANCE OF QUALITY AND SAFETY OF BEEF. *International Journal on Smart Sensing & Intelligent Systems*, *8*(1).