G305

Comparison of K-Means and Agglomerative Hierarchical Clustering Method with Principal Component Analysis in Research Document Analysis

Emi Susilowati
Department of Informatics Engineering
University of Muhammadiyah Jakarta
Central Jakarta 10510
emi_emt@yahoo.com

Dwi Handoko
Badan Pengkajian dan Penerapan Teknologi
Balai IPTEKnet (http://www.bppt.go.id)
Jl. M.H. Thamrin No. 8 Jakarta Pusat
dwihdk@gmail.com

***Abstract*—Clustering is one of text processing techniques which is widely known and has been intensively used in *data mining*. Deciding the appropriate clustering method based on particular case study will produce an optimal cluster. One of deciding techniques is comparing one clustering methods with another. In this study, K-Means Clustering and Agglomerative Hierarchical Clustering Combined with Principal Component Analysis are compared to analyze research documents. Principal Component Analysis is used to reduce data dimension, since there are abundant features, approximately 5803 features, extracted from the documents. The result of K-Means and Agglomerative Hierarchical Clustering with PCA comparison based on validation test shows that K-Means produced lowest Sum of Square Error (SSE) which is 10088.4313 with number of features is 100. It can be concluded that chosen methods, K-Means, for the number of class k=9 generated labels automatically, thus we can get the meaning for each chosen method. The extracted meanings become new valuable knowledge from the dataset of research documents. Based on 9 classes that has been mentioned, we can see research trends for food security category. As for several years later, the best theme on food security is still around expansion of production land and reduction of land loss, while the topic of research is related to test of multi-location expectation furrow and creation of superior varieties.**

Keywords—clustering, k-means, hierarchical agglomerative clustering, principal components analysis, food security;

I. INTRODUCTION

Science and technology will be a force for the advancement of Indonesia , and also become a source of pride in the nation's life , when the research and development of science and technology and science and technology utilization activities carried out in harmony and mutually reinforcing (National Research Agenda, 2010). Research is very important in determining whether or not a developed and developing countries . So what about the development of science and technology (science) in Indonesia ? " Science Indonesia has reached its nadir . Research institutions we already marginalized , a phenomenon which can be directly seen from the poor quality of research results . The scientific community and the general public have been shoulder to shoulder pressing research quality " (Terry Mart , Reuters May 8, 2006 edition) . Low science and research activities in Indonesia outline caused by several factors such as : the establishment of a national seed research theme is not based on the competencies of Human Resources (HR) ; lack of international collaboration ; less optimal performance reviewer team ; lack of raw determination to the research criteria good for all disciplines ; lack of government intervention in improving the quality of national research (TerryMart , http://staff.fisika.ui.ac.id/tmart/nadir.html , access October 8, 2011).

Research Incentive Program is a program funded by the Ministry of Research and Technology in building science and technology associated with the development of the National Innovation System . Ministry of Research and Technology along with the National Research Council (DRN) sets out a number of products targets to be achieved through Intensive Research Program . For the 2010 fiscal year provided 50 product targets are translated into 294 activities by giving priority to the seven (7) areas of focus of the development of science and technology as listed in the National Short-Term Development Plan (RPJPN) and the 2005-2025 National Medium Term Development Plan (RPJMN) 2010 -2014 , one of them is the field of food security . The number of research activities at the Research Incentive Program conducted by academics and researchers produce data in the form of course report documents the results of research in large numbers . In the research report documents containing the data : title research ; agencies ; budget ; field of research ; locations ; year ; abstract ; etc. . These data when administered using certain techniques can yield valuable information in the form of knowledge. Techniques or methods are known as data mining. Han, J. and Kamber , M. (2006:7) , Provide the definition of data mining as the process of discovering interesting knowledge from large amount of data stored in databases , data warehouses , or other information storage . Cluster techniques include techniques that are well known and widely used in data mining . The main purpose of the methods or techniques cluster is a grouping of a number of data or objects into clusters (groups) so that in each cluster will contain the data as closely as possible. Clustering is one of the unsupervised learning techniques which do not need to train

them or in other words there is no learning phase (Santosa , Budi . 2007:33) .

Clustering techniques can be used for document clustering research reports , so that the expected output of the clustering process can yield valuable information in the form of knowledge (knowledge) is needed as to identify trends or trends in any category of research focus areas of science and technology development , research themes what is in demand by the academics and researchers , institutions involved in the research , any location that is frequently used in research and much more valuable information that can be obtained from extracting data from documents of research reports . Selection of the appropriate cluster method will produce the optimal cluster . One way of selecting the appropriate cluster method is to perform a comparison between the cluster method to cluster the other methods . In this study discusses the results of the analysis report documents the results of a comparison study using a cluster of K - Means and Hierarchical Agglomerative Clustering with Principal Components Analysis . Use of Principal Component Analysis is used to reduce the dimension of data that has a large amount because featrure (word unique money) in this study is 5828 . To obtain a quality cluster can not be determined based on a subjective opinion , but must go through testing the validity of the cluster . In this study the validity of the clusters are used for the K - Means and Agglomerative Hierarchical Clustering is the Sum of Squared Error (SSE) . Expected from the comparison of the two methods can be seen right cluster method based on the determination of optimal clusters based on the validity of the cluster that has the value of Sum of Squared Error is the smallest that can be analyzed to document and interpret research reports Research Incentive Program based automatic label formed.

II. FOUNDATION FRAMEWORK FOR THINKING

According to (Chakrabarti et al., 2009:3), "Data mining is defined as the process of discovering patterns in the data. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The Data is present in substantial quantities Invariably ". In this definition submitted that as the process of data mining to find patterns in the data. Patterns discovered must be meaningful and bring some advantages, usually an economic advantage. Data is always there in large numbers.

The definition of data mining is also proposed Han, J. and Kamber, M. (2006:7), "Data mining is the process of discovering interesting knowledge from large amounts of the data stored in databases, data warehouses, or other information repositories". Data mining is the process of discovering interesting knowledge from large amount of data stored in databases, data warehouses, or other information storage.

A. Partitioning Method

K-Means is the most popular and very easy to understand in partitonal clustering algorithm. The main idea of the K-Means can be described in the following steps (Arriyani, 2010):

1. Determine the desired number of clusters k

2. Initialize k cluster center (centroid) by random / random

3. Place each data object to the cluster or nearby. The proximity of the two objects is determined by the distance. Distance used in the K-Means algorithm is Euclidean Distance.

4. Recalculate the cluster centers cluster membership now. Cluster center is the average (mean) of all the data or objects within a particular cluster. Back again to step 3 until no more objects moving clusters.

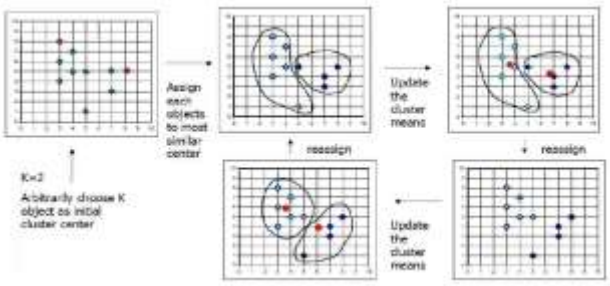

Fig. 1. Examples of K-Means Algorithm
Source: Han, J. and Kamber, M. (2001)

B. Hierarchical Method

Agglomerative hierarchical clustering algorithm to consider starting the process all the data as a set of initial clusters. Then clusters are combined in a larger group, this process continues over and over until you have the desired group. Agglomerative algorithms can be described in 3 steps (Chen, 2004):

1. Change object attribute into the distance matrix

2. Place each object as a cluster (if it has N objects, meaning have N clusters at first.

3. Repeat step two until the number of clusters is one, by the way:

   a. Combine the two (2) nearest cluster

   b. Update the value of the distance matrix

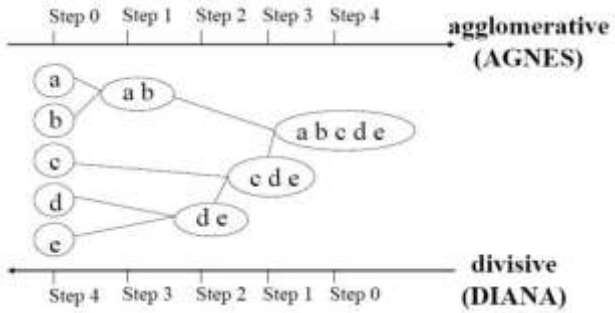

Fig. 2. Agglomerative and divisive Hierarchical Clustering Data Objects {a, b, c, d, e}

C. Principal Component Analysis

The dimensions of the data is determined by the number of features on the object. One method used to reduce the number of dimensions of the data is PCA (Principal Component Analysis). PCA procedure basically aims to simplify the observed variables by means of shrinking (reducing)

dimension. This is done by eliminating the correlation between the independent variables through the transformation of the independent variables to the origin of a new variable that is not correlated at all (Soemartini, 2008). With the help of matlab software, can use PCA procedure to reduce high correlated variables so as to assess which variables are truly worthy to be included in subsequent analyzes.

III. RESEARCH METHODOLOGY

This research was conducted in several stages in the framework described in Figure 3.1.

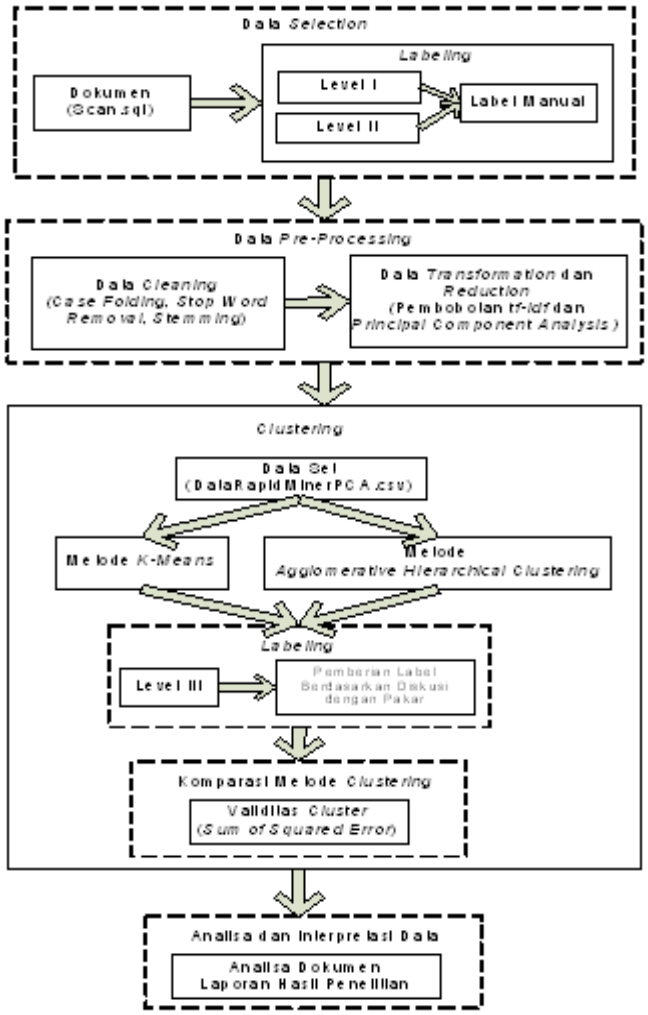

Fig. 3. Research Framework

A. Data Selection

In this study, the data used is document research reports Research Incentive Program in 2010 in food security, amounting to 391 documents in the form of structure query language (SQL) and then transformed to forms xml extension to facilitate the processing of data. The first step in data processing is a research report documents labeling or labeling manually, by the provision of food security categories for level I and level II for 5 subcategories. Sample report documents the results of research in the field of food security in 2010 is as follows:

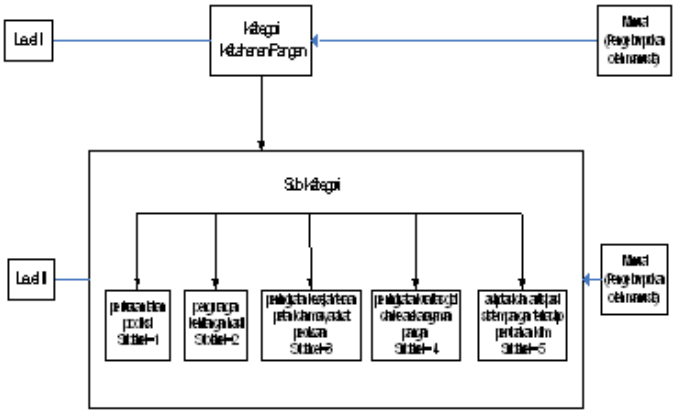

Fig. 4. Document Field Research Report on Food Security Source: Agency for the Assessment and Application of Technology (BPPT, 2010)

Fig. 5. On the Labeling Process Document Manually (Grouping by humans)

B. Data Preprocessing

To get the most out of data preprocessing stage used in the Java programming language and matlab. Data preprocessing stage is conducted in this study as follows:

1. Beheading words (tokens) in a document with the aim to facilitate the elimination of punctuation, numbers and words that are included in the stop word. Making all the words contained in the document to lowercase everything (case folding). Only the letter 'a' to 'z' that can be processed. The goal is to eliminate the duplication of the word.

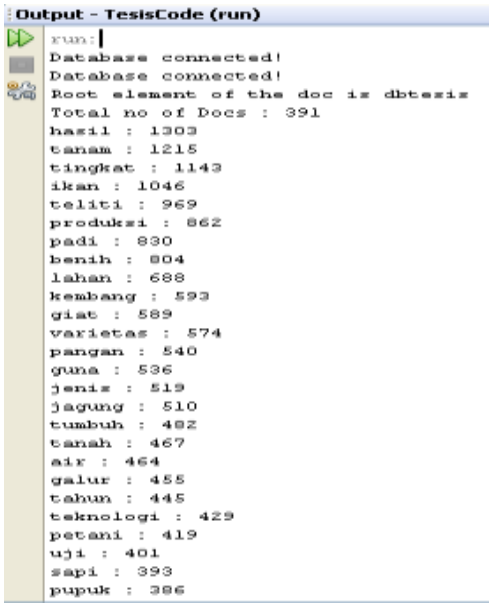

Fig. 6. Token process or decoding of words in a document

2. Making all the words contained in the document to lowercase everything (case folding). Only the letter 'a' to 'z' that can be processed. The goal is to eliminate the duplication of words.

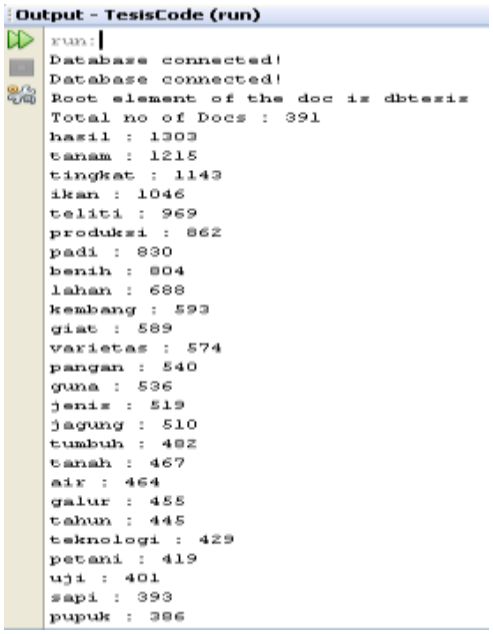

Fig. 7. Token process or decoding of words in a document

3. Do stop word or words that have no meaning but has a frequency of occurrence is very frequent in the document.

Fig. 8. Examples of words include Stop Word

4. Change all said additive into the base word (stemming)

id	term	stem
1	pengembangan	kembang
2	pembangkit	bangkit
3	peningkatan	tingkat
4	mendukung	dukung
5	pembuatan	buat
6	memanfaatkan	manfaat
7	meningkatkan	tingkat
8	pemakaian	pakai
10	kehandalan	handal
11	kesiapan	siap
12	kebijakan	bijak
13	mendorong	dorong
14	peralatan	alat
15	peranan	peran
16	pertumbuhan	tumbuh
17	ketahanan	tahan

Fig. 9. Examples of the results stemming the words in the document

5. Data transformation or data transformation is a process by which data is transformed into a form that can be in mining-right. The way that the object word (term) can in mining-it is a way of weighting the word using TF-IDF (Term Frequency - Inverse Document Frequency). Here is an example of a data matrix containing the results of weighting features 19 objects and 5 attributes.

TABLE I. SAMPLE DATA MATRIX FEATURE WEIGHTING RESULTS

-1.3847	2.481939	1.667287
-1.25787	-1.60353	-0.6839
-0.83293	0.789304	1.675255
-1.09657	-1.46045	1.89896
-1.50955	-1.72425	0.276499
-1.47843	-1.7297	-1.28111
-1.00914	-1.47069	1.303258
-0.71909	0.911802	2.255076
-1.44674	-1.64476	1.595621
-1.53786	-1.73888	1.264755
-0.59976	-1.30967	0.852068
-0.35283	7.387291	1.855891
-1.56131	-1.79753	0.042046
-1.30633	-1.68313	0.95556
-1.52221	0.670898	2.081172

In the table above can be explained from the data that the report documents the results of research into a form that can be transformed in ¬-right-mining to be understood by the data mining tool. Transformation of the data in this study using feature weighting tf-idf approach (term frequency - inverse document frequency).

6. Data reduction or data reduction is a way to reduce the number of data dimensions. The dimensions of the data is determined by the number of features on the object. One method used to reduce the number of dimensions of the data is PCA (Principal Component Analysis). PCA procedure basically aims to simplify the observed variables by means of shrinking (reducing) dimension. This is done by eliminating the correlation between the independent variables through the transformation of the independent variables to the origin of a new variable that is not correlated at all (Soemartini, 2008). With the help of matlab software, can use PCA procedure to reduce the independent variables are highly correlated so as to assess which variables are truly worthy to be included in subsequent analyzes.

TABLE II. EXAMPLE OF CALCULATION RESULTS FEATURE WEIGHTING USING PCA (DATA MATRIX CONTAINING 15 3 OBJECTS AND ATTRIBUTES)

-1.3847	2.481939	1.667287
-1.25787	-1.60353	-0.6839
-0.83293	0.789304	1.675255
-1.09657	-1.46045	1.89896
-1.50955	-1.72425	0.276499
-1.47843	-1.7297	-1.28111
-1.00914	-1.47069	1.303258
-0.71909	0.911802	2.255076
-1.44674	-1.64476	1.595621
-1.53786	-1.73888	1.264755
-0.59976	-1.30967	0.852068
-0.35283	7.387291	1.855891
-1.56131	-1.79753	0.042046
-1.30633	-1.68313	0.95556
-1.52221	0.670898	2.081172

The table above can be explained as follows, from the original data feature weighting totaling 391 objects (m) and 5803 attributes / term (n) is then calculated using a PCA with the help of matlab software, feature weighting calculation results obtained by PCA with the object number 391 and the number of attributes 100.

C. Clustering

This study used comparison of two methods, namely K-Means clustering and Agglomerative Hierarchical Clustering by Principal Components Analysis which serves to reduce the dimension of data that has a large amount. To measure the distance between two points proximity euclidean distance is used as the data dimension suitable for many of the most popular and used to calculate the dissimilarity, dissimilarity calculating formula is shown in equation 1.

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{in}-x_{jn})^2}, \quad (1)$$

D. Cluster Validity

The validity of the cluster is the procedure to assess the quality of clustering results and find a good strategy for cluster-specific applications. The validity of this cluster aims to find an optimal cluster and can interpret the pattern of the resulting clusters. In this study the validity of clusters used is Sum of Squared Error (SSE). The smaller the value of SSE will be better. Here is a formula of the SSE:

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2 \quad (2)$$

In equation 2 Sum of Squared Error (SSE) is used to determine the clustering result is better, if the initialization centroidnya different. In this study to find the optimal cluster from the comparison method K-Means and Agglomerative Hierarchical Clustering is used on the calculation of the smallest SSE. Techniques to minimize the value of SSE is to increase the value of k (Atastina, ITT Telkom).

E. Clustering Comparison Method

In this research, comparative method of K-Means and Hierarchical Agglomerative Clustering based on the results of the cluster, the processing time and the value of Sum of Squared Error (SSE).

F. Data Analysis and Interpretation

Based on the selection method of the best clusters were compared between the 2 methods, namely: K-Means and Agglomerative Hierarchical Clustering clusters are obtained optimal results. The resulting optimal clusters can be used to analyze the report documents the results of research by way of interpreting the data in the cluster formed. The results at this stage of the analysis and interpretation of this data can find new information that is interesting and valuable.

IV. RESULT AND DESCRIPTION

A. K-Means clustering

In the K-Means method, should be determined in advance the number of clusters desired. In this study, the number of clusters (k) that is used is k = 5, k = 7, k = 9. Here is the result of the clustering of each predefined number of clusters from the beginning of the K-Means method.

TABLE III. RESULTS OF K-MEANS CLUSTERING (K = 5)

label / cluster	Perluasan Lahan Produksi	Pengurangan Kehilangan Hasil	Peningkatan Kesejahteraan Petani dan Masyarakat Pedesaan	Peningkatan Kualitas Gizi dan Keanekaragaman Pangan	Adaptasi dan Antisipasi Sistem Pangan Terhadap Perubahan Iklim	Label	Anggota Cluster
0	13	6	4	2	0	Pengembangan Budidaya dan Teknologi	25
1	1	0	0	0	0	Evaluasi Sosial Ekonomi	1
2	22	1	0	0	0	Uji Multilokasi Galur Harapan	23
3	24	0	1	1	0	Pengkajian Penerapan Kebutuhan dan Sistem Penyediaan Ket	26
4	11	2	0	0	0	Pengkajian Teknologi Reproduksi	13
5	223	36	22	9	10	Pemberdayaan Keterse Unggul	382
6	1	0	0	0	0	Kajian Keterkaitan Produksi, Perdagangan dan Konsumsi	1
						Jumlah Data Set	381

TABLE IV. RESULTS OF K-MEANS CLUSTERING (K = 7)

label / cluster	Perluasan Lahan Produksi	Pengurangan Kehilangan Hasil	Peningkatan Kesejahteraan Petani dan Masyarakat Pedesaan	Peningkatan Kualitas Gizi dan Keanekaragaman Pangan	Adaptasi dan Antisipasi Sistem Pangan Terhadap Perubahan Iklim	Label	Anggota Cluster
0	13	6	4	2	0	Pengembangan Budidaya dan Teknologi	25
1	1	0	0	0	0	Evaluasi Sosial Ekonomi	1
2	21	0	0	0	0	Uji Multilokasi Galur Harapan	21
3	0	1	0	0	0	Assesed Usaha Tani	1
4	15	3	1	0	0	Pengkajian Teknologi Reproduksi	19
5	23	0	1	1	0	Pengkajian Penerapan Kebutuhan dan Sistem Penyediaan Kebutuhan	25
6	3	0	0	0	0	Reproduksi dan Mobilitas Sapi Kembar	3
7	220	35	21	9	10	Pemberdayaan Keterse Unggul	295
8	1	0	0	0	0	Kajian Keterkaitan Produksi, Perdagangan dan Konsumsi	1
						Jumlah Data Set	381

TABLE V. RESULTS OF K-MEANS CLUSTERING (K = 9)

label / cluster	Perluasan Lahan Produksi	Pengurangan Kehilangan Hasil	Peningkatan Kesejahteraan Petani dan Masyarakat Pedesaan	Peningkatan Kualitas Gizi dan Keanekaragaman Pangan	Adaptasi dan Antisipasi Sistem Pangan Terhadap Perubahan Iklim	Label	Anggota Cluster
0	296	44	27	12	10	Pemberdayaan Keterse Unggul, Pengkajian Penerapan Kebutuhan dan Teknologi Reproduksi	389
1	1	0	0	0	0	Evaluasi Sosial Ekonomi	1
2	1	0	0	0	0	Kajian Keterkaitan Produksi, Perdagangan dan Konsumsi	1
						Jumlah Data Set	381

B. Agglomerative Clustering with hierrachical Clustering (AHC)

AHC clustering method, not determined in advance the number of clusters desired, since the number of clusters obtained from the cutting (cut off) the right dendogram based on the largest gap of each stage and then specify the number of clusters are formed based on the results of the cut-off.

TABLE VI. AGGLOMERATIVE HIERARCHICAL CLUSTERING RESULTS CLUSTERING (AHC)

label / cluster	Perluasan Lahan Produksi	Pengurangan Kehilangan Hasil	Peningkatan Kesejahteraan Petani dan Masyarakat Pedesaan	Peningkatan Kualitas Gizi dan Keanekaragaman Pangan	Adaptasi dan Antisipasi Sistem Pangan Terhadap Perubahan Iklim	Label	Anggota Cluster
0	296	44	27	12	10	Pemberdayaan Keterse Unggul, Pengkajian Penerapan Kebutuhan dan Teknologi Reproduksi	389
1	1	0	0	0	0	Evaluasi Sosial Ekonomi	1
2	1	0	0	0	0	Kajian Keterkaitan Produksi, Perdagangan dan Konsumsi	1
						Jumlah Data Set	381

AHC clustering method, not determined in advance the number of clusters desired, since the number of clusters obtained from the cutting (cut off) the right dendogram based on the largest gap of each stage and then specify the number of clusters are formed based on the results of the cut-off.

C. Comparison of value Sum of Squared Error (SSE)

From the results of running the Sum of Squared Error (SSE) of 20 times running by considering the number of features from the range of 100 up to 400 objects. SSE is obtained as follows:

TABLE VII. COMPARISON OF THE RESULTS OF SSE VALUE

Cluster	*K-Means*			AHC
Objek	k=5	k=7	k=9	
100	**10754.928**	**10184.21173**	**10088.431**	**10849.704**
150	12555.056	11997.8994	11683.569	12506.799
200	13859	13280.9117	13070.446	13724.117
250	14614.999	14330.6002	14237.411	14623.203
300	15363.228	15045.5535	14903.604	15252.708
350	15736.686	15494.1683	15319.227	15689.444
400	15935.178	15695.6816	15518.633	15888.821
Rata2	14117.011	13718.4323	**13545.903**	**14076.4**

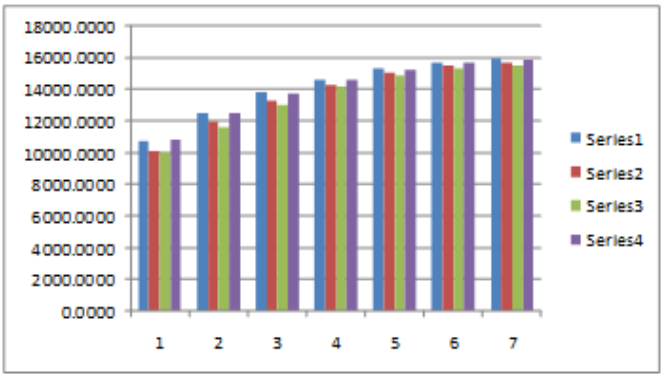

Fig. 10. Comparison of the results of SSE value for the K-Means method (k = 5, k = 7, k = 9) and AHC Methods

Based on Table VII. and figure 10 on the results of the comparison value of SSE, SSE average value for K-Means is the smallest number of clusters 13545.9030 at k = 9 and Agglomerative Hierarchical Clustering (AHC) has an average value of the SSE at 14076.3995. This indicates that the average value of the best (smallest) value of SSE is the number of K-Means cluster of nine (9). This means that the method of K-Means cluster with the number nine (9) to find members of each cluster are more similar than the AHC method.

D. Selected Results Clustering Analysis Method

Clustering results are used to make the analysis and interpretation of the data is the result of clustering using K-Means elected dengann number of clusters k = 9. This is because the value of the SSE of the K-Means method is smaller than the AHC method, which means that the error in the method of K-Means clustering smaller.

TABLE VIII. RESULT CLUSTERING METHOD CHOSEN BY K-MEANS (K = 9)

label cluster	Perluasan Lahan Produksi	Pengurangan Kehilangan Hasil	Peningkatan Kesejahteraan Petani dan Masyarakat	Peningkatan Kualitas Gizi dan Keanekaragaman	Adaptasi dan Antisipasi Sistem Pangan Terhadap	Anggota Cluster	Label
1	13	6	4	2	0	25	Pengembangan Budidaya dan Teknologi
2	1	0	0	0	0	1	Evaluasi Sosial Ekonomi
3	21	0	0	0	0	21	Uji Multilokasi Galur Harapan
4	0	1	0	0	0	1	Asuransi Usaha Tani
5	15	3	1	0	0	19	Pengkajian Teknologi Reproduksi
6	23	0	1	1	0	25	Pengkajian Pemetaan Kebutuhan dan Sistem Penyediaan Kebutuhan
7	3	0	0	0	0	3	Reproduksi dan Mobilitas Sapi Kembar
8	220	35	21	9	10	295	Pembentukan Varietas Unggul
9	1	0	0	0	0	1	Kajian Keterkaitan Produksi, Perdagangan dan Konsumsi
					Jumlah Data Sets	391	

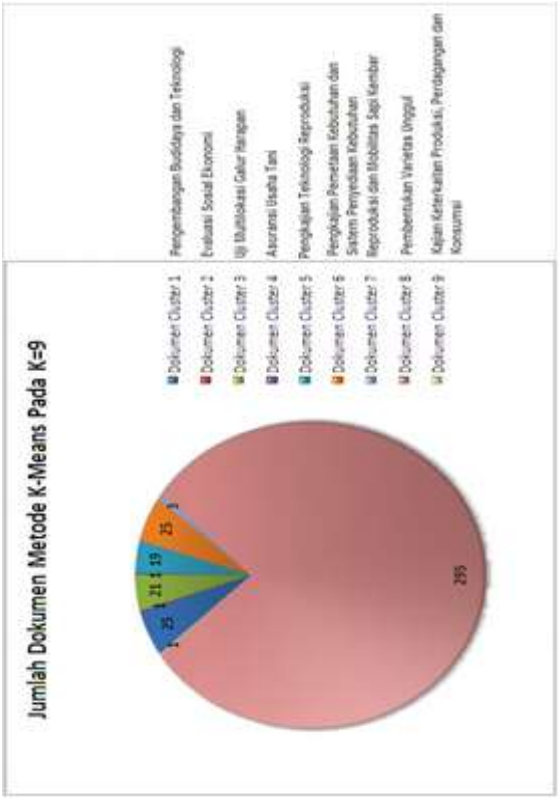

Fig. 11. Number of documents per-Cluster

Explanations for Table VIII. and Figure 11 are follows:

1. For Cluster 1 is a class where researchers concentrated on the theme of food security research subcategories expansion of production, reduction of yield loss, improving the welfare of farmers and communities, and improving the nutritional quality and diversity. Making more research topics range of Aquaculture Development and Technology.
2. For Cluster 2 is a classroom in which researchers concentrate on the theme of food security research subcategories expansion of production. Making research topics on Economic Social Evaluation. Small number of members can be considered as outliers or extreme data state that is different from the rules in general. It can be concluded that in 2010, the research topics that discuss the socio-economic evaluation is very rare.
3. For cluster 3 is the class in which the researchers concentrated on the theme of food security research subcategories expansion of production. Making all research topics range from about Strains Hope multilocation trials.
4. For cluster 4 is a class in which the researchers concentrated on the theme of food security research subcategories reduction in yield loss. Making research topics about Farm Insurance. Small number of members can be considered as outliers or extreme data state that is different from the rules in general. It can be concluded that in 2010, the research topics that discuss Farm Insurance is very rare.
5. For cluster 4 is a class in which the researchers concentrated on the theme of food security research subcategories reduction in yield loss. Making research topics about Farm Insurance. Small number of members can be considered as outliers or extreme data state that is different from the rules in general. It can be concluded that in 2010, the research topics that discuss Farm Insurance is very rare.
6. For cluster 6 is a class where researchers concentrated on the theme of food security research subcategories expansion of production, improving the welfare of farmers and communities and increased nutritional quality and diversity. Making more research topics range of Mapping Needs Assessment and Provision System Requirements.
7. For cluster 7 is a class where researchers concentrated on the theme of food security research subcategories expansion of production. Making all research topics range of Reproduction and Mobility Cattle Twins.
8. For cluster 8 is a class where researchers concentrated on the theme of food security research subcategories expansion of production, reduction of yield loss, improving the welfare of farmers and communities, and improving the nutritional quality and diversity, and Adaptation and Food Systems Against Climate Anticipation. Making more research topics range from about Development Establishment Varieties.
9. For cluster 9 is the class in which the researchers concentrated on the theme of food security research subcategories expansion of production. Making research topics on Assessment of Trade Linkages, Production and Consumption. Small number of members can be considered as outliers or extreme data state that is different from the rules in general. It can be concluded that in 2010, the research topics that discuss Linkage Study of Trade, Production and Consumption extremely rare.

E. Research Implications

The results of the comparison method of K-Means and Hierarchical Agglomerative Clustering with PCA based validation test smallest SSE values obtained can be concluded that the chosen method is the method of K-Means for k = 9, the label is automatically generated to obtain meaning for each selected cluster method, meaning that produced is knowledge (knowledge) is a valuable new set of documents from a set of data research reports. Of the nine (9) classes can be seen above the trend of the future for research on food security category of research that still revolves around the theme of excellent research on Land Expansion of Production and Reduction of Loss results, the research topics related to strain Hope multilocation trials and Formation of Varieties.

V. Conclusion and Recommendation

A. Selected Results Clustering Analysis Method

From the results of research conducted on the data selection to analysis and interpretation of data for analysis report documents the results of research using the comparison method of k-means and agglomerative hierarchical clustering with principal component analysis it was concluded that data mining techniques can be used to analyze the report documents the results of research based clustering procedure because of the comparison method of K - means and Hierarchical Agglomerative clustering with PCA based validation test values obtained Sum of Squared Error (SSE) smallest , this means that the chosen method is the method of K - means for k = 9 is a method with optimal clustering results where cluster labeling method that has elected to have a sense of knowledge (knowledge) is a valuable new set of documents from a set of data research reports . Of the nine (9) classes that have been formed by clustering automatically by the program and labeling (labeling) involving food security experts can be seen in the future research trends for categories of food security research that excellent research theme still revolves around the expansion of production and loss reduction the results of the research topics related to the formation or assembly of high-yielding varieties or clones or varieties of seeds .

B. Selected Results Clustering Analysis Method

For subsequent studies that combine partitioning methods and hierarchical methods. Calculations performed prior to the K-Means method and proceed with the calculation method of Agglomerative Hierarchical Clustering. To reduce the number of dimensions for the large amounts of data can use the Singular Value Decomposition or Principal Component Analysis. To get the optimal cluster can use two (2) test the validity as well as to increase the level of accuracy of the results of the cluster, such as the F-Measure and Purity.

Acknowledgment

The authors would like to thank Department Head of Informatics Engineering, and Dean of Faculty of Engineering University of Muhammadiyah Jakarta that have given a chance to participate AEMT conference as speaker.

References

[1] Chris Ding dan Xiaofeng He. *Principal Component Analysis and Effective K-Means Clustering*

[2] Chapman, et all., 2000. CRISP-DM 1.0 : *Step by Step Data Mining Guide*, CRISP-DM *Consortium*

[3] Han, J. dan Kamber, M (2001) . Data Mining : *Concepts and Techniques First Edition*, Morgan Kaufmann

[4] Han, J. dan Kamber, M (2001) . Data Mining : *Concepts and Techniques Second Edition*, Morgan Kaufmann

[5] Kusrini dan Luthfi, Emha Taufiq (2009). Algoritma Data Mining, Andi Publisher, Yogyakarta

[6] Santosa, Budi. (2007). Teknik Pemanfaatan Bisnis Dengan Data *Mining*, Graha Ilmu

[7] Rizky, Wibisono, Kusnendar. (2009). Perbandingan *Clustering Based on Frequent Word Sequence* dan *K-Means* Untuk pengelompokan Dokumen Berbahasa Indonesia

[8] Umran, Munzir dan Abidin, Taufik Fuadi (2009). Pengelompokkan Dokumen Menggunakan *K-Means* dan *Singular Value Decomposition* : Studi Kasus Menggunakan Blog

[9] Tajunisha dan Saravanan (2010). *An Increased Performance of Clustering High Dimensional Data Using Principal Component Analysis*