DATA MINING APPLICATION TO FORM A STUDY GROUP USING CLUSTERING METHOD AT SMPN 19 BANDUNG

Andre Catur Prasetyo

Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
E-mail : andrecatur86@gmail.com

ABSTRACT

SMPN 19 Bandung is one of the junior high school in Bandung. As schools that become the favorite choice, SMPN 19 Bandung must maintain the quality of their education. 9th grade students are usually required to follow the stabilization or try out. In addition to stabilization, school should undertake a study group formation. The problem is that the school usually form groups based solely on the order of attendance list. Which will cause excellent students are on the same group with students who left behind in a subject and it is worried the excellent students will be bored because the material provided is understood to be repeated so that students who left behind can catch up.

In data mining, there are methods that can be used to divide the data into groups based on the similarity of data, the method is clustering. In Clustering are also several methods that can be used, one of which is an algorithm Agglomerative Hierarchical Clustering (AHC) using single linkage algorithm. AHC process by using a single linkage algorithm begins with determining the number of groups will be formed, considers all data as clusters, calculating the distance matrix, looking for two adjacent cluster then combine them, then repeat step 3 until the remaining is number of clusters to be formed.

Based on testing results it can be concluded that the Application Study Group Former Application has been helping the curriculum to form study group that is appropriate based on the similarity of students value in each group.

Keywords: data mining, clustering, agglomerative hierarchical clustering, single linkage, distance matrix

1. INTRODUCTION

Learning is a requirement for junior high school students. When sitting in 9th grade student must establish their ability because they will face the graduation exam, the National Examination. SMPN 19 Bandung is one of the junior high school in Bandung, precisely at street Sadang Luhur XI, Bandung. As school that become the favorite choice, SMPN 19 Bandung must maintain their education quality.

Based on observations and interviews that have been conducted with Mrs. Denti as Vice Principal Curriculum in SMPN 19 Bandung, all students in 9th grade in the second semester are required to attend additional classes so-called Stabilization and try out test in order to prepare for the National Exam. But, there are still students who are left behind in certain subjects because every student has a different potential from each other. The student grade should be able to cross the minimum completeness criteria line in all subjects to graduate. In addition to stabilization, school should undertake a study group formation.

School usually form groups based solely on the order of attendance list. If the grouping is based on the attendance list, it is probable students with high grades in some subjects will be in the same group with students who left behind in these subjects. This will cause the students with high grades will be bored because the material is already understood to be repeated to help students who left behind understand the material.

To be able to catch up in a subject, study groups should be set up in accordance with the similarity of students grades. That will ease the teacher to identify where the difficulties experienced student on each study group was formed.

Therefore to resolve the above problem, data mining as one method to explore the knowledge can be used to analyze data of student's academic. One method that can be used is clustering method. Clustering method is a technique in data mining to categorize data that have similar characteristics with other data. By using clustering, the study group that resulting will consist of students who have a similarity in the value of academic subjects.

While objectives to be achieved in this research is to help the curriculum to form study groups based on similarity of student scores.

2. RESEARCH CONTENTS

2.1. Theoretical Basis

The theoretical basis on this thesis will explain the theories relating to the Data Mining Application

To Form A Study Group Using Clustering Method At Smpn 19 Bandung.

2.1.1. Data Mining

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. [6].

Data mining is an amalgamation of data knowledge and statistical analysis that was developed in the business knowledge or a process using statistical techniques, mathematics, artificial intelligence / imitation and machine-learning to extract and identify useful information for related knowledge from large databases [7].

The tasks of data mining comprise about knowledge extraction, archaeological data, the exploration in the processing of data patterns and harvest information. All of this activity is performed automatically and allows for rapid discovery even by non-programmers [7].

2.1.2. Phases of Data Mining

Data mining has many names or other terms, such as mining knowledge from the data, knowledge extraction, data / pattern analysis, data archeology and data dredging.

Many people who use the popular term against data mining, Knowledge Discovery from Data, or KDD. Data mining is only seen as an important step in a process of knowledge discovery. Phases of the data mining process is depicted in Figure 1.

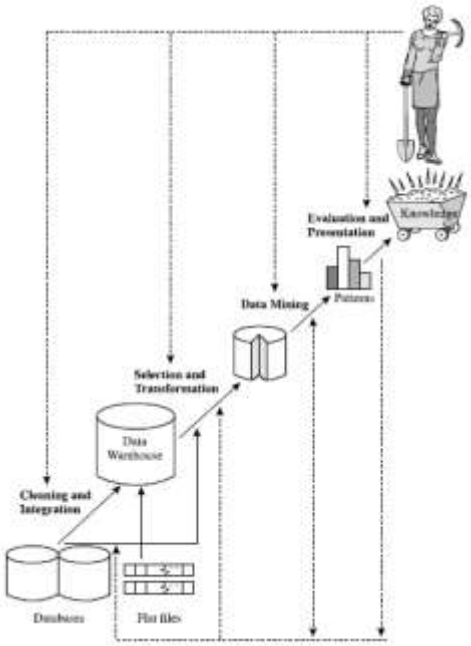


Figure 1 Data Mining on KDD Process [8]

The process of data mining in Figure 1 can be explained as follows:

1. Data Selection

Selection of a new data set of operational data needs to be done before the stage of extracting information in the KDD (Knowledge Discovery from Data) begins. Data from the selection that will be used for data mining process, is stored in a separate file from operational databases.

2. Preprocessing/Cleaning

Before the data mining process can be implemented, it needs cleaning process on the data to be the focus of KDD. Cleaning process encompass removing duplicates the data, check the data inconsistent, and correct errors in the data, such as printing errors.

3. Transformation

Coding is the process of transformation on the data that has been selected, so that data corresponding to the data mining process. KDD coding process in the creative process and is very dependent on the type or pattern of information to be searched in the database.

4. Data Mining

Data mining is the process of looking for patterns in the data or interesting information that is selected by using techniques or methods. Teknin, methods or algorithms in data mining is very varied. Selection of the appropriate method or algorithm relies heavily on the goals and overall KDD process.

5. Interpretation / Evaluation

Pattern information generated from the data mining process needs to be presented in a form that easily understood by stakeholders. This stage is part of the KDD process called interpretation. This phase includes checking whether a pattern or information found contradicted by the facts or hypotheses that existed before.

2.1.3. Data Mining Method

To complete the data mining research, there is a standard that can be used to complete the data mining research, namely Cross - Industry Standard of Data Mining (CRISP-DM). CRISP-DM is a comprehensive data mining methodologies and process models that give anyone (from beginner to expert data mining) is a complete blueprint to perform data mining projects [12].

Here are the stages of the methodology CRISP-DM [12]:

1. Business Undertanding

In the early stages of understanding of business focus to the understanding of business objectives, and then transform this knowledge into a data mining problem and then develop the initial plan designed to achieve the goal..

Business understanding phase divided into several steps, inter alia:

a. Determine Business Objectives

This stage aims to understand the business objectives to be achieved.

b. Determine data mining goals

The purpose of this phase is to determine the criteria for success of data mining.

2. Data Understanding

Data understanding phase is a phase to understand the data relating to the research conducted. Stage of understanding the data is divided into the following steps:

a. Collect Initial Data

In the collect initial data is used to collect data to be used in data mining.

b. Describe Data

Data which obtained from the collect initial data, described in describe data phase. In the describe data will be described the format of the data, quantities of data, the number of records and fields in each table, etc.

c. Explore Data

Analyzing the data which obtained previously include a descriptive statistical analysis and data visualization.

1. Descriptive statistical analysis

Descriptive statistical analysis used to help achieve the goal of data mining. Descriptive statistical analysis with regard to finding the values below:

a) Minimum value
b) Maximum value
c) Mean value using the following equation [10]:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

d) The standard deviation value using the following equation [10]:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (2)$$

2. Data Evaluation

In the data evaluation phase will examine the data by describing how the existing data, a way to see if there is a missing value, or outliers

a) Missing Value
b) Outliers

Outliers are data objects that deviate significantly from other objects [10]. The method can be used to detect outliers based on statistical techniques by using threshold. To obtain the threshold value can use equation 3.

Upper threshold = mean + (2 * standard deviation) (3)

Lower threshold = mean - (2 * standard deviation) (3)

3. Data Preparation

Data preparation phase include all activities forming the final dataset or data to be used in the modeling. There are several stages in the preparation phase of this data, inter alia:

a. Select Data

Specify the data to be used for the analysis based on specific criteria. Selection of this data includes the atribut selection or row selection.

b. Clean Data

Data cleaning phase aims to clean up or eliminate the data generated in the phase of evaluation of previous data.

c. Initial Data Preparation

This phase is used to store data that have been and do the cleaning to used on modeling phase.

4. Modeling

In this modeling phase perform the election of the model and apply the appropriate model. Modeling steps are as follows:

a. Select Modeling technique

This phase is used to select the appropriate modeling techniques to problems and goals to be achieved.

b. Build Model

In this phase described the modeling techniques selected.

c. Assess Model

In this phase the chosen model is applied to the test case.

5. Evaluation

Before the final phase, a model that has been made will be evaluated and reviewed in advance to ascertain whether the model can achieve the goal or not. There are several stages in the evaluation phase, which are:

a. Evaluate Result

This phase evaluating which model is the result closer to business objectives.

b. Review Proses

Conduct a thorough review of the involvement to ensure their data mining an important factor or ignored task.

c. Determine Next Steps

6. Development

The development phase is the phase of implementation for the development of applications in the form of representation of the knowledge that has been obtained so that it can be used by users.

2.1.4. Data Preprocessing

There are several things to note in the data quality of data mining. There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability. Data preprocessing can improve the quality of data, thus increasing the accuracy and efficiency of data mining results [13]. Some of the activities undertaken preprocessing of data in this study are as follows.

a) Missing Values

The existence of empty values may cause errors in the data. Empty values are usually caused

by human error (human error) when entering data. Therefore, empty data needs to be handled with the appropriate method. In this study, each object containing an empty data will be deleted.

b) Noisy Data

Noisy data is a random error or variance in a measured variable. Outliers can be represented as noise. As well as empty data, noise can also affect the quality of the data, therefore noisy data needs to be removed or eliminated in order to produce quality model. One way that can be done to clean up the noisy data is to perform smoothing. One smoothing technique that can be done is binning method.

Binning method divides the data set into multiple partitions or bin. Starting with any sort on an attribute value. Then the data that already sorted divided into multiple partitions or bin that have the same frequency (equal-frequency partitioning). There are two ways of smoothing the binning method, ie smoothing by bin means and smoothing by bin boundaries.

In smoothing by bin means performed by changing each value in the bin by the mean of the bin. While in smoothing by bin boundaries every value in the bin was changed to the lower limit (minimum) and lower limit (maximum) on each bin.

2.1.5. Clustering

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. According to Jiawei Han and Micheline Kamber, cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a formof data compression. [8].

Clustering is also called data segmentation in multiple applications for clustering divide large data sets into groups according to similarity. In machine learning, clustering is one example of unsupervised learning. Unlike classification, clustering does not rely on a predefined classes and training classes. Therefore, clustering is a form of learning by observation.

According to Santosa, basically clustering is a method to search and classify data that have similar characteristics (similarity) between one data with another. [9].

Jiawei Han and Micheline Kamber said that in general, the primary method of clustering grouped into some category is as follows [8].

1. Partition Methods
2. Hierarchical Methods
3. Density-based Methods
4. Grid-based Methods
5. Model-based Methods
6. Clustering high-dimensional data
7. Constraint-based clustering

The important things in the clustering is to measure the similarity and dissimilarity data to be grouped. The next will form the matrix within which explain the similarity between one data with other data or the distance between the data with the other data. The equation used to calculate the distance between the two data are Euclidean Distance. Here is the equation Euclidean Distance. [5]

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2} \quad (4)$$

d(i,j) distance between i to j
xi1 = value of the 1st attribute to x_i
xj1 = value of the 1st attribute to x_j
n = number of attribute

2.1.6. Agglomerative Hierarchical Clustering

In hierarchical clustering, cluster which are in the upper level (intermediate level) from another cluster can be obtained by combining the two clusters that are below the level. The overall result of hierarchical clustering can be graphically depicted as a tree, called dendogram.

In agglomerative hierarchical clustering, hierarchical clustering process starting from the clusters that have points of individuals who are in the lowest level. At every step, the merger of a cluster with other clusters, where clusters are coupled to each other or have a high degree of similarity [10].

One of the algorithms used in agglomerative hierarchical clustering is single linkage. Single linkage occurs when groups merged based on the closest distance from one cluster to another cluster.

The steps from Agglomerative Hierarchical Clustering with single linkage algorithm is as follows [11]:

1. Specifies k as the number of clusters to be formed.
2. Each data is regarded as a cluster. If n = number of data and c = number of clusters, then c = n.
3. Calculating the distance matrix / similarity / dissimilarity between clusters (eg. using euclidean distance that contained in equation (4))
4. Find two clusters that have the smallest distance between clusters and merge them, so that c = c-1.
5. If c> k, repeat steps c and d until c = k.

In the single linkage, the closeness between the two clusters is determined by the distance of the nearest (smallest) of two different clusters. In this way it will be starting from the respective data as a cluster, then look for the nearest neighbor by using a single linkage to combine the two next cluster until it becomes a cluster.

Here's a single linkage equation [10]:

$d_{(UV)W} = \min\{d_{UW}, d_{VW}\}$ (2.5)

d_{UW}= distance between cluster u with cluster w

d_{VW}= distance between cluster v with cluster w

2.2. Problem Analysis

Based on observations and interviews with the vice principal of curriculum, obtained the following results:

1. School usually form groups based solely on the order of attendance list.
2. if clustering performed based on attendance list, the division of the group members would not be evenly distributed. Students with high grades in certain subjects can be in the same group with students who have difficulty in these subjects.
3. With the the uneven division of the group, will lead to students with high grades will be bored because the material provided is repeated to help students who left behind can catch up.

2.3. Business Understanding

Business understanding phase is an early stage in the framework of the CRISP-DM. This phase focuses on understanding the purpose and requirements from a business standpoint SMPN 19 Bandung. There are several tasks at this stage, inter alia:

2.3.1. Determine Business Objectives

SMPN 19 Bandung has a business purpose that is capable of guiding the student to pass in each subject by passing the minimum completeness criteria boundary. An important factor to business goals SMPN 19 Bandung achieved namely by organize programs that can improve a student's academic potential, one example is the forming study groups according to the similarity value of students.

2.3.2. Determine Data Mining Goals

1. Data mining objectives

The purpose on the application of data mining student data in this research is to form study groups based on midterms grades of student who have the same characteristics.

2. Successful data mining criteria

Criteria for success against this research that if able to classify the data based on midterms grades of student who have the same characteristics. Size of group quality will be measured by comparing the grouping process that is running with the grouping process of clustering techniques in data mining.

2.4. Data Understanding

Data understanding phase is the secend phase of the framework of the CRISP-DM. There are several tasks at this stage is:

2.4.1. Collect Initial Data

The data used in this research is odd midterm grades of student from class 9E, 9F, and 9G Year 2015. The data that used was the pure student grades from midterm results without remedial.

Student grades data to be used in the form of file format *.xls or *.xlsx. In addition the data used in this study only the data that is in the responsibility of SMPN 19 Bandung.

2.4.2. Describe Data

The data used has 7 attributes, 3 first attribute profiles of students, while 4 other attributes explain the value of Middle Semester Exam. The amount of data contained in the data record value of as much as 107.

attributes contained in the odd midterms grades data from class 9E, 9F, and 9G year 2015 described in Table 1 below.

Table 1 Data Attribute Explanation

No	Atribut	Tipe Data	Keterangan
1.	No Induk	String	A student identification code that is unique
2.	Nama	String	A student name
3.	Jenis Kelamin	String	Gender of student
4.	Nilai UTS B.Ind	Numeric	Midterm grades of Indonesian Language subject
5.	Nilai UTS B.Ing	Numeric	Midterm grades of English Language subject
6	Nilai UTS Mat	Numeric	Midterm grades of Mathematics subject
7.	Nilai UTS IPA	Numeric	Midterm grades of Science subject

2.4.3. Explore Data

Explore data phase can help the purpose of data mining, in this explore data comprise descriptive analysis and data quality evaluation. Data that used in this phase is midterms grades of student on class 9E. Here are the results of data exploration against the data students grades are:

1. Descriptive analysis

This descriptive analysis performed against attribute Nilai UTS B.Ind, Nilai UTS B.Ing, Nilai UTS Mat, dan Nilai UTS IPA. Analysis carried out on these attributes because clustering attributes which will be done based on midterms grades of student.

In this phase will be performed analysis against midterms grades of student from class 9E

2015, the analysis performed on each attribute midterms grades of student. The following analysis will be done, among others:

a. Minimum value
   Here the minimum value of midterm grades on each subject:
   Min Nilai UTS B.Ind = 50
   Min Nilai UTS B.Ing = 68
   Min Nilai UTS Mat = 20
   Min Nilai UTS IPA = 70
b. Maximum value
   Here the maximum value of midterm grades on each subject:
   Max Nilai UTS B.Ind = 80
   Max Nilai UTS B.Ing = 88
   Max Nilai UTS Mat = 100
   Max Nilai UTS IPA = 100
c. Mean Value
   To get mean value can use equation (1). Then using equation (1) obtained the following results:
   1)Mean value of nilai UTS B.Ind

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{2470}{36} = 68.61$$

   2)Mean value of nilai UTS B.Ing

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{2850}{36} = 79.17$$

   3)Mean value of nilai UTS Mat

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{2575}{36} = 71.53$$

   4)Mean value of nilai UTS IPA

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \frac{3101}{36} = 86.14$$

d. Standard Deviation
   Using equation (2) obtained the following results:
   Table 2 Descriptif Analysis Results

No	Subjects	Min value	Max value	Mean value	Standard Deviation
1	B. Indonesia	50	80	68.61	8.05
2	B. Inggris	68	88	79.17	4.90
3	Matematika	20	100	71.53	24.57
4	IPA	73	100	86.14	7.75

2. Data Quality Evaluation

Data quality evaluation performed against attribute Nilai UTS B.Ind, Nilai UTS B.Ing, Nilai UTS Mat, dan Nilai UTS IPA. Here the result from data quality evaluation:

a. There is a missing value in the data class 9E on attributes no induk. Missing value on no_induk attributes will be changed to be PINDAHAN1, PINDAHAN2, etc.
b. The value of the data will be seen whether outlier or not, and the following is a way that can to do to looking for outliers using equation (3). Then gained the upper threshold and lower threshold as can be seen in Table 3.

Tabel 1 Batas Atas dan Batas Bawah Outlier

No	Subject	Upper Threshold	Lower Threshold
1	B. Indonesia	84.71	52.51
2	B. Inggris	88.97	69.37
3	Matematika	120.67	22.39
4	IPA	101.94	70.34

2.5. Data Preparation

This phase includes all activities required to construct the final dataset (data that will be incorporated into the modeling tools) from the initial raw data, there are some tasks in this phase, inter alia:

2.5.1. Select Data

Select data have the task of includes attribute and row election. The choosen attribute. The selected attributes to make the process of grouping (clustering) among others, No Induk, Nilai UTS B.Ind, Nilai UTS B.Ing, Nilai UTS Mat, Nilai UTS IPA.

2.5.2. Clean Data

In this phase carried out data cleaning process, based on quality verification obtained result as follow:

1. Missing value found on attribute no induk. Missing value on attribute no induk will be replaced by PINDAHAN1, PINDAHAN2, etc.
2. value which is outlier will still be processed, to perform student clustering will be used algorith that are resistant to outlier existance, that is with perform smoothing process against data using binning method.

2.5.3. Initial Data Preparation

In the process of preparing the initial data will be prepared the data to be used in the modeling phase. Data that will be used on modeling phase is

grades data of class 9E year 2015 that has been smoothed.

2.6. Modeling

The next phase of the framework of the CRISP-DM is modeling, this phase has the task of selecting modeling techniques, model making and model testing.

2.6.1. Modeling Technique

Modeling techniques are used in accordance with the original purpose of which is to form study groups. The model will be tested is Agglomerative Hierarchical Clustering (AHC).

2.6.2. Model Making

This phase explains the mechanism of the model to be used and the initial assumptions of the model. Model Agglomerative hierarchical clustering has the following step:

1. Specifies k as the number of clusters to be formed.
2. Each data is regarded as a cluster. If n = number of data and c = number of clusters, then c = n.
3. Calculating the distance matrix / similarity / dissimilarity between clusters (eg. using euclidean distance that contained in equation (4))
4. Find two clusters that have the smallest distance between clusters and merge them, so that c = c-1.
5. If c> k, repeat steps c and d until c = k.

2.6.3. Model Testing Analysis

The data will be used to perform this grouping is midterm grades of student data from class 9E year 2015. The amount of data used is as much as 36 records.

The case will be tested using AHC algorithm is as follows:

1. Specifies the number of cluster to be formed

The number of study groups to be set up is as much as three groups based on the number of groups are commonly formed is three. If k = number of clusters to be formed, then k = 3.

2. Each data is regarded as a cluster

Initialize initial cluster is about 36 because data(n) = cluster(c), then calculate the distance between the cluster with the existing cluster using Euclidean Distance equation (Equation 4).

3. Calculate the distance between each data

In this phase, each cluster will calculated the distance with himself and the distance to the other cluster. Here is the process of calculating distaance between clusters using Euclidean Distance equation (equation 4).

$$d(1,1) = \sqrt{(72-72)^2 + (79-79)^2 + (95-95)^2 + (93-93)^2} = 0$$
$$d(1,2) = \sqrt{(72-55)^2 + (79-76)^2 + (95-30)^2 + (93-80)^2} = 68.50$$
$$d(1,3) = \sqrt{(72-67)^2 + (79-79)^2 + (95-95)^2 + (93-91)^2} = 5.39$$
$$\vdots$$
$$d(36,36) = \sqrt{(72-72)^2 + (71-71)^2 + (81-81)^2 + (73-73)^2} = 0$$

With the same way, that is using Euclidean Distance equation performed against all data. By calculating Euclidean Distance, obtained distance matrix.

4. Looking for two adjacent clusters

In this phase will be calculated the distance between two clusters using a single linkage method (equation 5), with this method will look for the closest distance of the two clusters.

1st Iteration:

$$d_{(1,2)3} = \min\{d_{(1,3)}, d_{(2,3)}\} = \min\{68.50, 5.39\} = 5.39$$

until

$$d_{(u,v)n} = \min\{d_{(u,n)}, d_{(v,n)}\}$$

With single linkage method, the closest distance from each iteration is done will combine two clusters, so that the number of iterations that will be generated from the data and the number of 36 clusters to be set up as many as k = 3 clusters will produce as much as 33 iterations. If the cluster (n, n) and euclidean distance = 0, then the value of the euclidean distance is not expressed as the smallest and continue the search. Suppose cluster (1, 1), cluster (2.2), cluster (3,3), and so on that have a value euclidean distance = 0. From the first iteration of the distance that is closest cluster of 13 to 23 with a distance value is 2:00 , then both these clusters are combined into one cluster and the number of newly formed cluster as many as 35 clusters.

5. If c > k, repeat step 3 and 4

Step 3 and 4 performed repeatedly and will stop when the number of formed clusters (c) = number of clusters to be formed (k) or c =k =3. To get three cluster, in this case need to be done as many as 33 iteration.

From the grouping process by using AHC obtained the following results:

a. By using the cluster number as many as three clusters of the amount of data that the data used were 36, 7 data are in cluster 1, 22, the data is in cluster 2 and 7 data in cluster 3.
b. By using a single linkage clustering method will tend to congregate in one cluster because each iteration performed looking for the smallest value of each cluster, and the value is used to comparison with another cluster.
c. Cluster 1 and Cluster 3 has an average value which is quite low in Indonesian language and mathematics subjects.
d. Cluster 2 has an average value which is quite low in Indonesian language subjects.

2.6. Use Case Diagram

Use case or use case diagram is a model for the behavior of the system being designed. Use case diagram contained on the system will be built consists of a ten actors and use case.



Figure 2 Use Case Diagram Study Group Former System SMPN 19 Bandung

2.7. Class Diagram

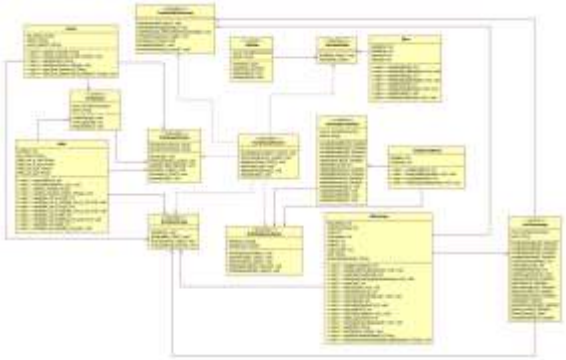


Figure 1 Class Diagram

2.8. Relationship Scheme

Relationship scheme is a set of relationships between two or more tables in the database system. Here is a relationship scheme in the study groups former system in SMPN 19 Bandung:



Figure 2 Relationship Scheme

2.9. Menu Structure Design

The design of the menu structure on the applications that will be built only for one user. Here is the menu structure design application to be built.

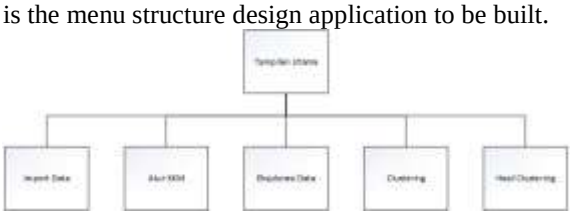


Figure 3 Menu Structure

2.10. Interface Design

The design interface is the design of the display system to be built which serves as a communication medium that is used as a means of interaction between the system with the users. The system built is expected to provide a display that is easy to understand and easy to use by the user. The design of the interface on the system built is as follows.

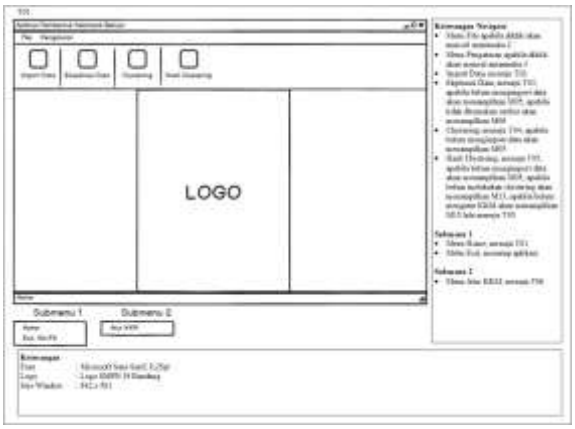


Figure 4 Main Menu Interface Design

2.11. Semantic Networks

Semantic Network is connectedness of a menu navigation from one page to the other page. Semantic network of the system to be built can be seen in Figure 7.

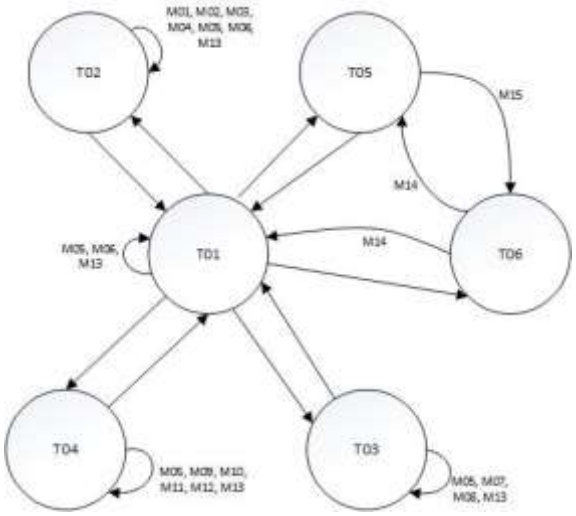


Figure 5 Semantic Network

2.12. System Testing

System testing performed on the system functionality to assess whether applications that are built already meet the criteria or there is a mismatch in functionality as contained in the system design. The test are used to test the software to be built is a black box testing methods.

Test plan which will do is by testing the application to be built using black box and beta testing. Black box testing focus to functionality testing of software to be built.

1. Black Box Testing

Black box testing focus to functional requirement of software to be built.

Tests carried out by testing each process for possible mistakes.

Based on the result of testing by tes sample case that have been done, can concluded that all function running well and goes as expected.

2. Beta Testing

Beta testing is a test that is performed directly in the actual environment. Users give an assessment of the software that was built through the interview method. Based on the results of these interviews, it can be concluded whether the software is built in accordance with the purpose or not.

Interviews were conducted for software built with the staff part of the curriculum. This research was conducted at SMPN 19 Bandung in order to determine the extent to which software was built to be a solution and solving problems that have been described previously.

Based on the beta testing has been done before, it can be concluded that an application was built can help curriculum to the form appropriate study groups based on similarity of student grades.

3. CLOSING

In this section, describes the conclusions which contains the results obtained after the analysis, design, and implementation of the software design was built and has been developed as well as suggestions that will provide important records and possible improvements that need to be done for software development previously.

3.1. Conclusion

The results of the research conducted was concluded that the application was built help curriculum to form study group in accordance with the similarity of students grades.

3.2. Suggestion

There are some suggestions that can be done to develop study groups former applications, among others:

1. addition of features to set school year in order to improve the accuracy of clustering and allows the user to filter data of clustering results that will be used if the data were grouped mixture of different school year.
2. Development grouping can be developed further by using another algorithm so that the user can compare the results of the classification with Single Linkage algorithm.

BIBLIOGRAPHY

[1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, CRISP-DM 1.0, SPSS Inc, 2000.

[2] Kadir and Abdul, Pengenalan Sistem Informasi, Yogyakarta: ANDI, 2003.

[3] G. B. Davis, Management Information System: Conceptual Foundations, Structure, and Development, Auckllland: McGraw-Hill International Book Company, 1974, p. 32.

[4] J. G. Burch and F. R. Strater, Information System: Theory and Practice, California: Hamilton Publishing Company, 1974, p. 23.

[5] G. R. Terry, Office Management and Control, 4th ed., Ilinois: Richard D. Irwin Inc., 1962, p. 21.

[6] D. T. Larose, Discovering Knowledge in Data, Hoboken: WILEY-INTERSCIENCE, 2005.

[7] S. Kurniawan and T. Hidayat, "PENERAPAN DATA MINING DENGAN METODE INTERPOLASI UNTUK MEMPREDIKSI MINAT KONSUMEN ASURANSI (Studi Kasus Asuransi Metlife)," vol. 5, pp. 113-128, 2007.

[8] J. Han and M. Kamber, Data Mining: Concepts and Technique, 2nd ed., San Fransisco: MORGAN KAUFMANN, 2006.

[9] B. Santosa, Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis, Yogyakarta: Graha Ilmu, 2007.

[10] G. A. Pradnyana and N. A. S. ER, "Perancangan dan Implementasi Automated Document Integration dengan Menggunalan Algoritma Complete Linkage Aggglomerative Hierarchical Clustering," Jurnal Ilmu Komputer, vol. 5, no. 2, pp. 1-10, September 2012.
[11] R. Handoyo, R. R. M and S. M. Nasution, "Perbandingan Metode Clustering Menggunakan Metode Single Linkage dan K-Means Pada Pengelompokan Dokumen," JSM STMIK Mikroskil, vol. 15, no. 2, pp. 73-82, Oktober 2014.
[12] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data," Journal of Data Warehousing, vol. 5, no. 4, p. 14, 2000.
[13] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques 3rd Edition, San Fransisco: Morgan Kauffman.
[14] R. Miles and K. Hamilton, Learning UML 2.0, Sebastopol: O'Reilly Media Inc., 2006.