Use Clustering Data of Student High School for Placement in Personalization E-Learning on Higher Education

Purwono Hendrad, Harry Budi Santoso, Zainal A Hasibuan

Abstract*: Personalize the e-learning begins after students interact with the system by utilizing the functions and features to collect data and process it so that the resulting information from students who used to organize further activities. In another study, the educational background of the student (and types of SMA) also affects the success in education at the university. In this study developed a personalized e-learning design of the early, which is when the new students will interact with the system. The system will be a kind of student placement test. The case studies used subjects Program Building which is one of the core subjects in the study program Engineering Informatics. As the methods used Knowledge Data Discovery (KDD) using background data combined with a high school student math scores on the National Exam as an ingredient on the stage of Data Mining. This study will measure the extent of the student's educational background above can be used as a system of placement of students in personalized e-learning.

Index Terms*: high school background, data mining, placement, personalized e-learning.

I. INTRODUCTION

Highlight a section that you want to designate with a certain style, and then select the appropriate name on the style menu. The style will adjust your fonts and line spacing. Do not change the font sizes or line spacing to squeeze more text into a limited number of pages. Use italics for emphasis; do not underline.

Web's personalization nowadays develops follow with the development of internet access devices. Personalization of web makes the system could recognize and adapted with user behaviour. From the personalization of web inspired to personalization of e-learning who can recognize and adapted with student behaviour. Personalization of e-learning will adjust the learning process with the type of student on a learning plan. From diverse types of students are expected to achieve the standard of competence planned. The Indonesian high school has two types, high school, and vocational high school. General high school divide by two programs, exact and non-exact, vocational high school divide many programs, for this paper could explain with three groups, computer program, non-computer exact program and non-computer non-exact. All of them must take the national exam for passed who called 'Ujian Nasional'. The exam has five subjects, one of which is mathematic. Originally majoring in high school can be used as the basis for selection of new admissions[1]. On the private higher education, the filter of the student not strictly, the effect is so many kinds of student. Therefore Personalized e-learning became solution who could serve and be bridging of many student types. Before student studied in personalize e-learning, the variation of student background above would identified with education data mining (EDM). EDM use on student identification as academic prediction and failure [2] [3] [4]. The purpose of this paper to cluster of students and compare with the score of the first quiz. Result from this could be recommendation for placement of student on personalized e-learning.

A. Theoretical Background

Student of high school Turkey filtered to be the student of higher education by the national exam, called nationwide entrance exam. In research at the department of Management Information System (MIS), Boğaziçi University did the data mining with clustering and find student profile influence by student high school background. Student from general high school more success than vocational high school on MIS subject [5]. On other research, forty-five percent Grade Point Average (GPA) of the student with good criteria originated from the student with good background on high school. That result get from the research of data clustering in a private college on Semarang central java Indonesia [6].

Personalization of learning is a strategy of learning who identify of student characteristic and then use for the learning process to be more effective learning. The form of personalization is the personalization of learning flow, personalization of interface and personalization of content. From technical use approach: data mining, semantic web, and predefined rule [7]. From the research of personalization e-learning, personalization begins after the student interacts with the system. The system recorded a log of the student, student activity on forum and assessment. That data will be learning behavior pattern and then process to characteristic layer with tripe characteristic factor [8] [9] [10]. From the figure 1 personalization at the lower layer, the upper layer is layer characteristic for identification of student's characteristic. Characteristic layer gets data from learning layer, especially from learning behavior. Personalization did after minimum student first interaction with the system.

Revised Version Manuscript Received on April 13, 2017.

Purwono Hendradi,**M.Kom** Teknik Informatika Universitas Muhammadiyah Magelang, Jl. Mayjend Bambang Soegeng KM 5 Magelang, Indonesia , E-mail: p_hendra@ummgl.ac.id

Harry Budi Santoso, M.Kom, Ph.D, Ilmu Komputer, Universitas Indonesia Depok Jawa Barat 16424, Indonesia, E-mail: harrybs@cs.ui.ac.id

Prof. Zainal A Hasibuan, Ilmu Komputer, Universitas Indonesia Depok Jawa Barat 16424, Indonesia, E-mail: zhasibua@cs.ui.ac.id

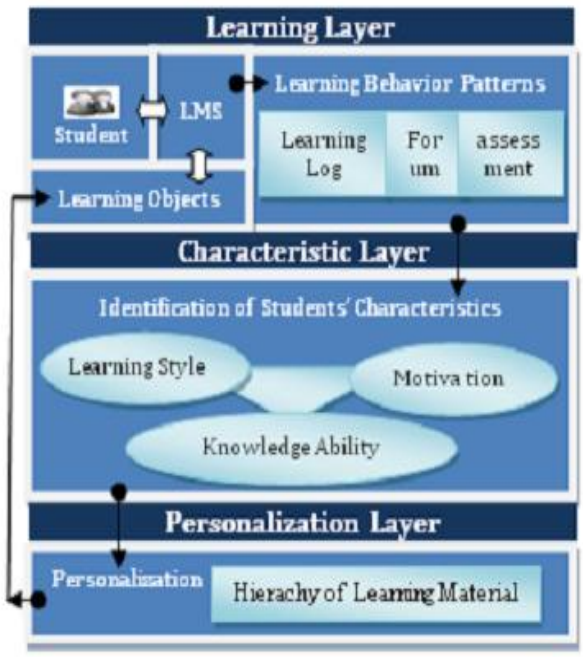

Fig 1. Triple-Characteristic Model (TCM) [8]

On placement research with a goal for prediction of student's performance on the university, use academic performance, technical skills, soft skills, training and project as measured parameter. The result is three separated kind, the student is ready and fulfill, student need to improved and students will face difficulty in completing his/her degree. The research use data mining, and the result will use system training that predicts the category of the student from a new database for next session [11].

On the other research for predicting the performance about the placement of final year students use a sum of difference for find pattern. Data gathered from a number of 50 students with 11 listed attributes were then processed for generating rules. The method sum of difference with simplification of data and then normalized could be efficient to predicts the placement of a student [3].

Inspiring from both of above research are student placement system could use student test records like academic performance, technical skills, soft skills, training, and project.

B. Knowledge Discovery in Databases [12]

Knowledge Discovery in Databases (KDD) is process looking for and identify the pattern of data become useful and easy to understand. This is the initial process for the data mining process. The following of the KDD schema:

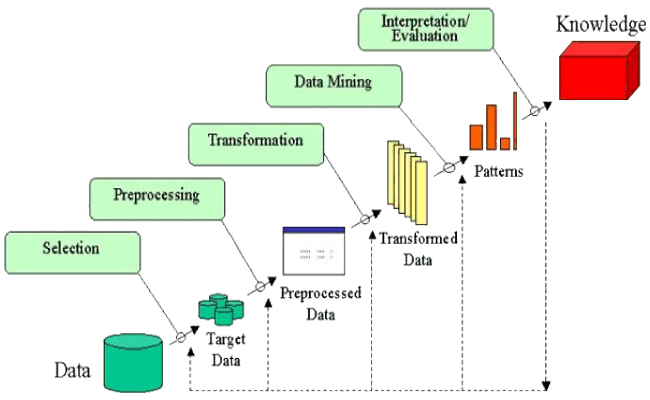

Fig 2. Step of Knowledge Discovery in Data (KDD) [12]

On figure 2 shown before data mining, beginning selection of target data, pre-processing and transformation of data. The result from data mining then did interpretation and evaluation to get knowledge.

For this research propose for placement system using data mining k-mean on personalization e-learning. High school data of student will cluster then compare with the result of the first quiz. This research take a sample from informatics department at the Muhammadiyah University of Magelang with program building's lecture on the first semester.

C. Data Mining

Data mining is combined with technique statistic, math, artificial intelligence and learning machine for to extract and to identify of information to useful knowledge from the big data [13]. Data mining based on task can be divided into some group, one of them is clustering as a method to search similarities characteristics between data each other. clustering is one of data mining method which has a nature unsupervised. [14]

D. Clustering K-Mean

K-Means clustering is basically a partitioning method for to analyze data as objects based on locations and distance between various input data points. Each cluster is characterized by its center point i.e. centroid.

from of several ways measure distance in k-mean, this research uses Euclidean distance. This below is the formula:

$$d(p-q) = \sqrt{(p1-q1)^2 + (p2-q2)^2} \qquad (1)$$

II. Material And Method

A. Material

The material of research is data origin of high school, the program in high school and score of mathematic on the national exam. Origin of high school is two type general high school (SMA) and vocational high school (SMK). From the origin of high school could be explained to program on high school.

On general high school program split exact program and non-exact program. Exact program is a program with a focus on natural sciences and the non-exact program is a program that focuses on social sciences. But on vocational high school consist of many programs. Because field of this research on computer program on higher education, then vocational high school program separated to three group. The first group is 'most supporting' like computer networking technician (TKJ), software engineering (RPL) and multimedia (MM), the second group is 'supporting' like all of another exact program like electrical, engine, machine etc. Then the last group is 'not support' like all nonexact program.

A score of mathematic on the national exam is a reliable value for measuring the ability math. Value will be categorized into 3 types; more than and equal 80 (>=80), between 60 to 79(60-79) and less than 60 (<60).

B. Method

Research method beginning clustering of student background data (high school) refer to KDD and method clustering k-mean.

The result will be compared with the score of the first quiz.

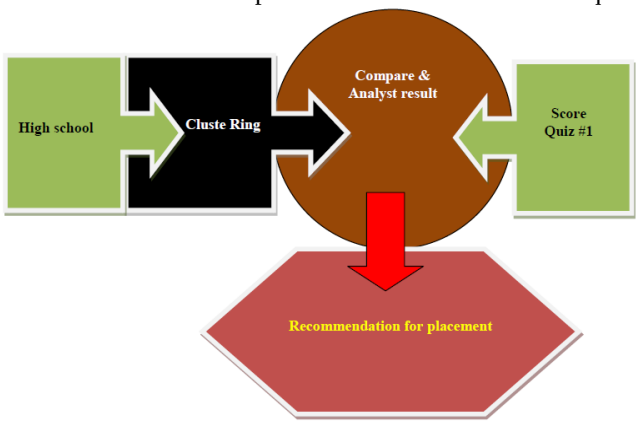

Fig 3. Research Method

Data of student background in high school using in this research refer to a purpose for the research. To get a recommendation into placement in personalization e-learning, the required data is a type of high school combined with program study in high school and mathematic score in the national exam.

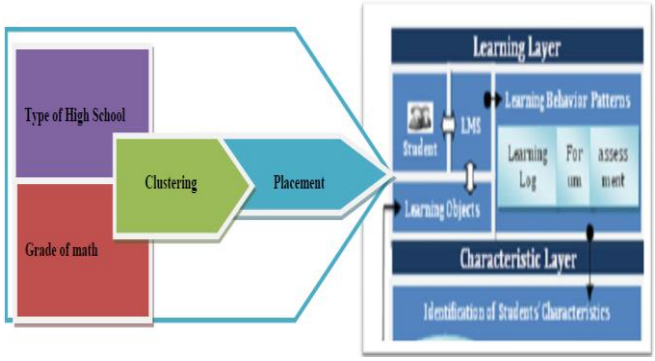

Fig 4. Placement system to Triple-Characteristic Model (TCM)

C. Algorithmic steps for K-Means Clustering [15]

1) Set K – To choose a number of desired clusters, K.
2) Initialization – To choose k starting points which are used as initial estimates of the cluster centroids. They are taken as the initial starting values.
3) Classification – To examine each point in the dataset and assign it to the cluster whose centroid is nearest to it.
4) Centroid calculation – When each point in the data set is assigned to a cluster, it is needed to recalculate the new k centroids.
5) Convergence criteria – The steps of (iii) and (iv) require to be repeated until no point changes its cluster assignment or until the centroids no longer move.

III. RESULTS AND DISCUSSION

The first step, collect the data background of students. This steps use questioner use Google form and filled in by a new student in informatics department. The question of the questioner are Name, NPM (student id number), date of birth, area, high school origin (asal Sekolah), program study at high school (jurusan), and mathematic score at the national exams. Total respondent is 76 students.

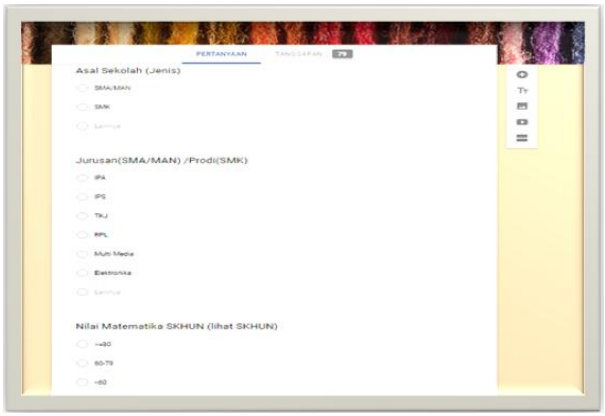

Fig 5. Google Form

Table 1. Raw Data

No	name	city	Origin high school	Program	Point of Math National Exam
1			SMK	TKJ	>=80
2			SMK	TKJ	>=80
3			SMA/MAN	IPA	<60
4			SMA/MAN	IPS	<60
			SMA/MAN	IPS	>=80
….	…….	…..	……	…..	….
76			SMK	sekrt	60-79

For the mathematic score of the national exam, student data is matched with the data range; >=80, 60-79 or <60, because the last time the student is inconsistent in writing down the value, some student fill 7 and other 75, that make confuse. From the data, processing with KDD:

A. Data Selection

From raw data, separated to personal data and nonpersonal data. NPM, Name, Date of birth and score at the national exam is personal data. Origin of High School and a program is nonpersonal data Origin of High School could separate two kinds, general high school (SMA) and vocational high school (SMK). For SMA could separate in two kinds, high school exact (IPA) and high school non-exact (IPS) and for SMK could be categorized into 3 types, 'most supporting', 'supported' and 'unsupported' with programming subject. 'Most supporting' program study on SMK are Networking Technician (TKJ), Software Engineering (RPL), Multi-Media (MM) and Electronic. then the study program as a 'supporter' is Automotive, Machine and engineering program others. For easy to understand, presented in the table below:

Table 2. High School and Program

School	Program	
SMA	Exact	
	Non Exact	IPS dan Bahasa
SMK	Computer	TKJ, MM, RPL,
	Exact	O to, Machine ,
	Non Exact	Accounting, Secretary etc

B. Data Transformation

Data transformation is done by changing to quantitative data based subjects 'program building'. High School (SMA) with program is exact sciences (IPA) is near or very support to subjects 'Program Building' (in higher education), so that the score is 3. Then all of data done transformed to value score, shown table below

Table 3. Tabel score of school

School	program	Score
SMA	Exact	3
	Non Exact	1
SMK	Computer	3
	Exact	2
	Non Exact	1

National Exam scores are transformed to be the quantity value also, shown table below:

Table 4. Table Score of National Exam Point

Value	Score
>=80	3
60-79	2
<60	1

The result of the above data would be processed to data clustering by the k-mean method.

C. Data Clustering

From the raw data (table 2) data transformed, the result can be explained in the table below:

Table 5. Transformation of Data

Origin high school	Program	Score	Point of Math / National Exam	Score
SMK	TKJ	3	>=80	3
SMK	TKJ	3	>=80	3
SMA/MAN	IPA	2	<60	1
SMA/MAN	IPS	1	<60	1
SMA/MAN	IPS	1	>=80	3
…….	…..	……	…..	…….

The scores of the High School and Math the result of the transformation of data are both of them in use for row material in Data Mining K-Mean. Then choose the data from the Table for determining the Centroid. From the table above choose three, the first data with the value score of the Program 3 and value score of Math exam 3 (3-3), then 2-2 and last 1-1. The data processed with the K-Mean formula, row by row use application Microsoft Excel.

```
=SQRT((D3-D$23)^2+(E3-E$23)^2)
```

Fig.6 Implementation k-mean to Ms Excel

The result of the Microsoft Excel formula shown in next three columns is distance to centroid 1, distance to centroid 2 and distance to centroid 3. From three columns, compared to search the minimum value to choose the Cluster to use the Microsoft Excel formula, shown on the picture bellow:

```
=IF(AND(G2<H2,G2<I2),1,IF(AND(H2<G2,H2<I2),2,3))
```

Fig.7 choose cluster in Ms Excel

Then the result would be shown in part of the Table in Table 6 until Table 9. The table below (Table 6) is the first iteration to K-Mean.

Table 6. Sample First Iteration to K-Mean

Distance to Centroid 1	Distance to Centroid 2	Distance to Centroid 3	cluster
2.828427	1.908215	0.717062	3
2.828427	1.908215	0.717062	3
1	0.449734	1.851276	1
0	1.003563	2.244085	1
2	1.803064	1.315891	2

Next process, result of the first iteration then be continued to the second iteration and third iteration. The result of iteration shown in the table below (table 7 and table 8):

Table 7.Sample 2nd iteration to K-Mean

Distance to Centroid 1	Distance to Centroid 2	Distance to Centroid 3	cluster
2.457807	1.564353	0.533333	3
2.457807	1.564353	0.533333	3
0.428571	0.917136	2.053723	1
0.571429	1.264403	2.480143	1
2.080031	1.400839	1.466667	2

Table 8. Sample 3rd Iteration To K-Mean

Distance to Centroid 1	Distance to Centroid 2	Distance to Centroid 3	cluster
2.828427	1.908215	0.717062	3
2.828427	1.908215	0.717062	3
1	0.449734	1.851276	2
0	1.003563	2.244085	1
2	1.803064	1.315891	3

Iterations in the data clustering performed at least 3 times and will stop when the centroid distance are relative remain (unchanged). On table 5, 6 and 7, is minimum iteration for K-mean. Iterated did from Transformation of data from table 5. After 3 times iteration, continue to 4th iteration, shown in the table below.

Table 9. Sample 4th iteration to k-mean

Distance to Centroid 1	Distance to Centroid 2	Distance to Centroid 3	cluster
2.513104	1.586388	0.717062	3
2.513104	1.586388	0.717062	3
1.108483	0.835817	1.851276	2
0.478261	1.245167	2.244085	1
1.521739	1.490077	1.315891	3

From table 9 shown data in cluster columns is same as data cluster columns in table 8. This describes the iteration on clustering is the enough and stop. Because data do not change and this is the maximum iteration. In this research, the iterations stop at fourth iteration because of the distance to centroid unchanged. This is consistent with the theory of K-Mean clustering that stopped when the iterating results have not changed compared to the previous iteration. After the fourth iteration, the result compares to the first score of the quiz in the Program Building subjects and this is the result. For this research, the quiz is classification into 3

groups, group >=80, 60-79 and <60 according to their academic qualifications.

The first quiz at the Program Building subjects is the quiz test of the student with the basic of programming command for the mathematic case like implemented to arithmetically mathematics to programming command. Student tries to implement to a simple script, the score depends on result and programming success.

The result of this quiz classified into 3 group as above and then data matched to the fourth iteration. Above Table (Table 10, 11 and 12) shown of the analyst of member each clustering result. Column quiz is the range of data first quiz on the 'Program Building' lecture, column QTY is the number of student data who get the score on that cluster. the Comparison column is the ratio of data per row on QTY with QTY accumulation, and the percentage column is rounding from the Comparison column and converted to percentage form.

Table 10. Analyst data cluster 1 with first quiz

Quiz	QTY	Comparison	Percentage
>=80	3	0.130435	13%
60-79	4	0.173913	17%
<60	16	0.695652	70%

Shown in table 10, is the result of analyst data in cluster 1 as a group of student with the low score. In table 70% (rounding of 0.695652) member from the student who get score of quiz less than 60. While the remaining 17% is the middle group and 13% is an upper group.

Table 11. Analyst data cluster 2 with first quiz

Quiz	QTY	Comparison	Percentage
>=80	11	0.366667	37%
60-79	5	0.166667	17%
<60	14	0.466667	46%

In the table above, result of analyst data in cluster 2 as a group of student with the middle score. Percentage of middle (60-79) just 17 % (rounding of 0.166667) smaller than quantity of low group (<60) and upper group (>=80) as 46% and 37%.

Table 12. Analyst data cluster 3 with first quiz

Quiz	QTY	Comparison	Percentage
>=80	13	0.565217	57%
60-79	1	0.043478	4%
<60	9	0.391304	39%

Then in table 12, the result of analyst data in cluster 3 as a group of student with the high group, dominated by the student with a score of quiz bigger and equal with 80 (>=80). But second place is a student with the score less than 60 (<60) is 39%. This is surprising because the student with a low score could get in the high cluster while middle score (60-79) just 4%. From three analyst table above, is easy to find low cluster student, as well as cluster high although has some surprise data. But for cluster of mid is difficult to find pattern, because all data of quiz get the same data distribution, nothing dominant.

In this research could find the pattern of the upper cluster and bottom cluster of a student but difficult to find the pattern of a middle cluster. Maybe sharpness of student data and quality of data influences the result.

Because position in middle cluster is not clear if compare with origin data of student. So the results could not be recommended for personalized e-learning students except when the original data develop for the better.

Possibility precision of data of clustering not enough to describe a character from background of student data. For future will try to break down data and add other data. For math score from the national exam will be presented as real data, not as a range. Then add to the score of recruitment test who the contain a measurement of verbal performance, quantitative performance and logic performance for recruitment test to be a student at the university. And the last add the level of the type of school (favorite or nor).

IV. CONCLUSION

The background of the student like origin and program study of the High School who combined with score mathematic of national exam could describe of student cluster for subjects in high education. This result can use for placement of the student to start study with an e-learning so that from the beginning the students get personalized treatment.

But if comparison the result and data origin are something to developed. In this research could find a pattern of the upper cluster and bottom cluster of the student but difficult to find a pattern of the middle cluster. Sharpness of student data and quality of data influence the result.

FUTURE WORK

First, more sharpened the origin data and combine the result of recruitment test with origin data of student and compare of the result with this research. Second,

Uses unsupervised clustering method for this case and compare the result with this result, like Fuzzy c-mean method.

REFERENCES

1. Yayah Karyanah, "Hubungan Asal Jurusan dengan Prestasi Belajar Mahasiswa Program Sstudi Ilmu Keperawatan Universitas Esa Unggul," Forum Ilmiah, vol. 12, no. 2, pp. 156-163, May 2015.
2. C., Romero, C., & Ventura Marquez-Vera, "Predicting School Failure Using Data Mining," in Proceedings of the 4th international conference on educational data mining, 2011, pp. 271– 275.
3. Swarnalatha P, D. Ganesh Gopal Ramanathan.L, "Mining Educational Data for Students' Placement Prediction using Sum of Difference Method," International Journal of Computer Applications, vol. 99, no. 18, pp. 36-39, August 2014.
4. Romero C. AND Ventura, "Educational Data mining: A Review of the State of the Art.," IEEE Transactions on Systems. Man, and Cybernetics., vol. 40, no. 6, pp. 601-618, 2010.
5. Bertan Y. Badur Osman N. Darcan, "Student Profiling on Academic on Academic Performance Using Cluster Analysis," Journal of e-Learning & Higher Education, vol. 2012, p. 8, 2012.
6. Narwati, "Pengelompokan Siswa Menggunakan Algoritma K-Means," Dinamika Informatika, pp. 12-16, 2010.
7. Zainal A. Hasibuan, Harry Budi Santoso Mira Suryani, "Personalisasi Konten Pembelajaran Berdasarkan Pendekatan Tipe Belajar Triple-Factor Dalam Student Centered E-Learning Environment," in KNSI , Makasar, 2014.
8. Zainal A Hasibuan Sfenrianto, "Triple Characteristic Model (TCM) in E-Learning System," Proceedings of 4th International Conference on Computer Science and Information Technology. IEEE, Chengdu, 2011.
9. Zainal A Hasibuan, Heru Suhartanto Sfenrianto, "An Automatic Approach for Identifying Triple-Factor in e-Learning Process," International Journal of Computer Theory and Engineering, vol. 5, no. 2, pp. 371-376, April 2013.
10. Zainal. A. Hasibuan and H. B. Santoso., "The Use of E-Learning towards New Learning Paradigm: Case Study Student Centered E-Learning Environment at Faculty of Computer Science - University of Indonesia," in Proc. IEEE International Conference on Advanced Learning Technologies (ICALT 05), Kaohsiung, Taiwan, 2005, pp. 1026-1030.
11. Rajan Vohra Praveen Rani, "Generating Placement Intelligence in Higher Education Using Data Mining," (IJCSIT) International Journal of Computer Science and Information Technologies, vol. Vol. 6, no. 3, pp. 2298-2302, May 2015.
12. Howard Hamilton. (2012, June) Howard J. Hamilton. [Online]. http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html
13. Daniel T Larose, Data Mining Methods and Models. Hoboken, New Jersey: Jhon Wiley & Sons, Inc, 2006.
14. Daniel T Larose, Discovering Knowledge in Data: An Introduction to Data Mining: John Willey & Sons. Inc, 2005.
15. T. Kanungo and D. M. Mount, "An Efficient K-means Clustering Algorithm: Analysis and Implementation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 35-39, 2002.