Journal of Theoretical and Applied Information Technology
15th January 2012. Vol. 35 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ANALYSIS AND IMPLEMENTATION OF ALGORITHM CLUSTERING AFFINITY PROPAGATION AND K-MEANS AT DATA STUDENT BASED ON GPA AND DURATION OF BACHELOR-THESIS COMPLETION

[1]R.REFIANTI, [2]A.B. MUTIARA, [3]A. JUARNA, [4]S.N. IKHSAN

[1]Asst.Prof., Faculty of Computer Science dan Information Technology, Gunadarma University, Indonesia

[2]Prof., Faculty of Computer Science dan Information Technology, Gunadarma University, Indonesia

[3]Assoc.Prof., Faculty of Computer Science dan Information Technology, Gunadarma University, Indonesia

[4]Alumni, Faculty of Computer Science dan Information Technology, Gunadarma University, Indonesia

E-mail: [1,2,3]{rina,amutiara,ajuarna}@staff.gunadarma.ac.id , [4]pincess.ucul@gmail.com

ABSTRACT

Effectiveness and accurate results from an algorithm has always been a basic reference for every step taken in the use and utilization of algorithm, which is expected to achieve optimal results both in quality and quantity. In order to realize the level of accuracy and effectiveness from the program, it would require an algorithm that can minimize error and faster in data processing rate compared with existing algorithm In this paper, we have compared two algorithms, namely Affinity Propagation and K-Means, at data student based on GPA and Duration of Bachelor-Thesis Completion. The results show that Affinity propagation gives the result of data cluster more accurate and effective than K-Means, it can be seen from the testing table which showing that the value of affinity propagation exemplar has not changed at all after five trials. While K-Means, gives values of its centroid are different after five trials. And at the data students itself, it show that there is a relationship between GPA and Duration of Bachelor-Thesis completion in Gunadarma University students, it can be seen from the results of data clustering, that is for student who have GPA above 3 to 4 have a tendency to finish their Bachelor-Thesis faster, which is less than 1 until 2 semesters. While other students who have GPA less than 3 have a longer time to finish their Bachelor-Thesis, within a period of 2 until more than 4 semesters.

Keywords: *Data Clustering, Affinity Propagation, K-Means, GPA, Thesis, Gunadarma University*

1. INTRODUCTION

1.1 Back Ground

The Role of Science and Technology particularly in the field of informatics has been increased in accordance with community needs, ranging from the education system, up to the daily work that can be facilitated by the lack of progress in this field of informatics. The progress is primarily supported by many developing new algorithms that can help a programmer to process data and create a new application of the algorithm.

Effectiveness and accurate results from an algorithm has always been a basic reference for every step taken in the use and utilization of this algorithm is expected to achieve optimal results both in quality and quantity. In order to realize the level of accuracy and effectiveness of the programs created, it would require an algorithm that can minimize error rates and faster data processing rate compared with existing algorithms.

Based on this situation, the journal is focused on the analysis for the used of two algorithm namely, Affinity propagation and K-Means, and the implementation of both algorithm, so they can be applied to data clustering Gunadarma University's students based on GPA and Duration of Bachelor-Thesis completion. Analysis performed to determine the best algorithm in terms of effectiveness an accuracy of data between Affinity propagation and K-Means as well as to determine whether or not the relationship between GPA with Duration of Bachelor-Thesis completion on Gunadarma University's students, using version

7.09 of MatLab Software in the process of testing both algorithm.

1.2 Goal

- Analyze two algorithm, Affinity Propagation and K-Means, in order to know which one is more accurate and effective in classifying the data
- Implement both of algorithms on data student based on GPA and Duration of Bachelor-Thesis completion at Gunadarma University, in order to know the connectivity between student's GPA and Duration of Bachelor-Thesis completion

2. THEORETICAL BASIS

2.1 Data Cluster

Clustering is a method of analyzing data, which often included as one of the methods of data mining, whose purpose is to group data with similar characteristics to the same area and data with different characteristics to other areas. There are several approaches used in developing the method of clustering.

Two main approaches are clustering with partition approach, and hierarchical approach. Clustering with partition approach is often referred to as partition or partition-based clustering to group the data by sifting through the data analyzed in the existing clusters. Clustering with the hierarchical approach is often called clustering hierarchical group data by creating a hierarchy in the form of a dendogram, in which data will be placed on a hierarchy similar to the adjacent and which are not that far apart on the hierarchy. Besides both approaches, there is also clustering with automatic mapping approach (Self-Organizing Map / SOM)

2.2 Algorithm

In mathematics and computing, the algorithm is a collection of command to solve a problem. These commands can be translated in stages from beginning to the end. These problems can be anything, with a record for every problem, and there are criteria for the initial conditions which have been done before running the algorithm. The algorithm will be always ends for all initial conditions that meet those criteria, in this case different from the heuristic.

Algorithms often have steps repetition (iteration) or require decisions (Boolean logic and comparison) until the job is completed. Design and analysis of the algorithm is a special part of computer science that studies the characteristics and performance of an algorithm in solving problem, regardless of the implementation of an algorithm. In this part of the discipline, an algorithm learned by abstractly, regardless of the computer system or programming language used.

Different algorithms can be applied to a problem with the same criteria. The complexity of an algorithm is a measure of how much computation is required by the algorithm to solve the problem. Informally, an algorithm which can solve the problems in a short time has a low complexity, while the algorithm that takes a long time to resolve problem has a high complexity.

2.3 Affinity Propagation

Affinity propagation is known in computer science as a message-passing algorithm, where each item will be grouped (as sender) sends a message to all other items (as recipient) in order to inform the relative attractiveness of each item of recipient to the sender. Each recipient then responds to all senders with a reply informing each sender of its availability to associate with the sender, given the attractiveness messages that it has received from all other senders. Senders absorb the information, and reply to the recipient with messages informing each recipient of the recipient's revised relative attractiveness to the sender, given the availability messages it has received from all recipients. The message-passing procedure proceeds until a consensus is reached on the best associate for each item, considering relative attractiveness and availability. The best associate of each item is that item's exemplar, and all items sharing the same exemplar are in the same cluster. Essentially, the algorithm simulates conversation in a gathering of people, where each in conversation with all other seeks to identify his or her best representative for some function.

Affinity propagation is an algorithm that is simultaneously considers all of data points as possible exemplar (center point) where each message is sent to reflect the latest interest which is owned by each data point to be able to select another data points as their exemplar. In other words, affinity propagation is also known as a cluster algorithm that implements message-passing to all data points. Through this message-passing process, an algorithm tests the possibility of all data points to become the center of cluster (exemplar). So that, every data point has the same opportunity to become an exemplar.

Generally, an algorithm operates in three matrices, they are: matrix similarity (*s*), matrix of responsibility (*r*), and availability matrix (*a*). At the beginning of algorithm is a process to cluster by divides data based on the similarity measure, so data will be easier to be characterized by their similarities. Similarity matrix formed by eliminating the distance between items. The distance is calculated by adding the squares of the differences between item's variables. Similarity *(S(i,k))* presented how well data point with index *k* corresponding to become an exemplar for data point *i*. Then, the algorithm is continuing with the process to determine how well the suitability of data point to become an exemplar for other data points, where the process is known as responsibility. Responsibility *(R(i,k))* is the process where message is sent from data point *i* to the candidate exemplar of data point *k*. Final stage is availability process, which is illustrate how right of data points to choose another data point as their exemplar, where the availability *(A(i,k))*is send a message from candidate exemplar *k* to data point *i*.

2.4 K-Means

K-Means is a method of non-hierarchical data cluster which is trying to partition data into one or more cluster. This method is partitioning data into cluster which has similar characteristics are gathered in the same cluster. K-Means also called as repeatedly clustering algorithm. K-Means algorithm starts with a random selection of *k*, where *k* is the number of clusters to be formed. Then set the values of *k* randomly, for a while that values will be a center of cluster or commonly called centroid, mean, or means. Calculate the distance of every data to each centroid using Euclidean formula to find the closest distance of each data with the centroid. Then classify each data based on proximity to the centroid. The migration of data point will be always happen until the centroid does not change again (stable).

K-Means is a method of analyzing data or methods that perform data mining modeling process without supervision (unsupervised) and it is one method that grouping data with the partition system. K-Means method is trying to classify data into several groups, where data in one group have the same characteristics with each other and have different characteristics from existing data in the other group. In other words, this algorithm seeks to minimize the variation among the data in one cluster and maximize the variation with data on other cluster.

The following is procedures in performing optimization using K-Means:

- Step 1: Determine the number of clusters
- Step 2: Allocate data into clusters randomly
- Step 3: Calculate the centroid/ average of data contained in each cluster
- Step 4: Allocate each data to the closest centroid
- Step 5: Go back to step 3, if data is migrates to another cluster or when values of centroid changing (not stable yet).

3. DISCUSSION

3.1 Problem Analysis

Analysis of the issues to be addressed is about the comparison between affinity propagation algorithm and K-Means inform of implemented on the relationship between two variables, GPA and Duration of Bachelor-Thesis completion at Gunadarma University. In order to see the relationship between those variables and also to know an algorithm that is more effective and accurate in classifying data to produce the most optimal cluster.

In a previous cluster algorithms such K-Means, has a weakness in determining the initial exemplar (initial value of the center point), where its initial exemplar were obtained randomly, so it can't determined exactly the initial exemplar which will provide the best cluster results. If the initial exemplar were close of good value, then the cluster will be maximal. But, if the initial approached the value of exemplar are less good, then the cluster will also be less than the maximum. So, in the K-Means algorithm can be occur fairly high error rate. Likewise, with the form of cluster, must be determined in advance, because this algorithm can not determine automatically the number of clusters can be generated.

While, in the affinity propagation, initial exemplar were no longer needed because all data points will be tested as candidates exemplar. Since all data points are likely to be an exemplar then a good or bad cluster results will not depend anymore on a good or bad initial exemplar provided.

3.2 Problem Solving

To resolve the issues that exist in K-Means, can be done by analyzing and comparing it with affinity propagation algorithm. Troubleshooting on affinity propagation consisted of several stages of problem resolution.

Based on the workflow stages of problem solving on affinity propagation above, it was found that the algorithm does not require an initial exemplar because all data points will be tested as a candidate exemplar repeatedly on the responsibility and availability process to determine whether data points are more suited to be the exemplar or member of the exemplar. So, the algorithm will produce low error rate when compared with K-Means algorithm.

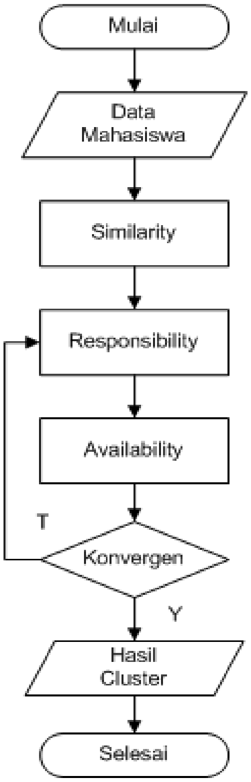


Figure 3.1. Flowchart of affinity propagation

4. TEST AND RESULT

4.1 Test

Fig. 4.1 and Fig.4.2 show the results of implementing the both algorithms.

In the testing stage, there are 5 times test of cluster on both algorithm using the same data. Data in used are as many as 50 data. Table 4.1 shows the results of Affinity Propagation testing on the data. After testing the cluster in affinity propagation, then also carried out cluster test in *K*-means algorithm. The results are shown on Table 4.2.

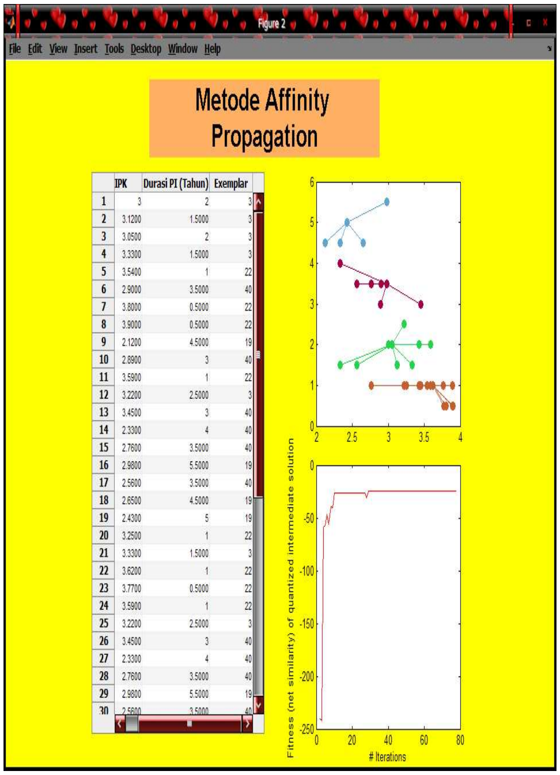
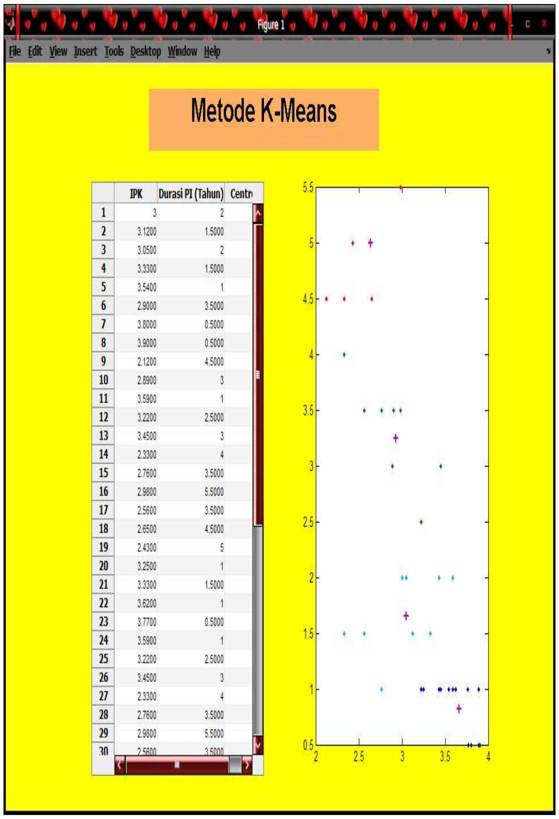


Figure 2
File Edit View Insert Tools Desktop Window Help
Metode Affinity Propagation
6
5
4
3
2
1
0
2
2.5
3
3.5
4
0
-50
-100
-150
-200
-250
0
20
40
60
80
Fitness (net similarity) of quantized intermediate solution
# Iterations

	IPK	Durasi PI (Tahun)	Exemplar
1	3	2	3
2	3.1200	1.5000	3
3	3.0500	2	3
4	3.3300	1.5000	3
5	3.5400	1	22
6	2.9000	3.5000	40
7	3.8000	0.5000	22
8	3.9000	0.5000	22
9	2.1200	4.5000	19
10	2.8900	3	40
11	3.5900	1	22
12	3.2200	2.5000	3
13	3.4500	3	40
14	2.3300	4	40
15	2.7600	3.5000	40
16	2.9800	5.5000	19
17	2.5600	3.5000	40
18	2.6500	4.5000	19
19	2.4300	5	19
20	3.2500	1	22
21	3.3300	1.5000	3
22	3.6200	1	22
23	3.7700	0.5000	22
24	3.5900	1	22
25	3.2200	2.5000	3
26	3.4500	3	40
27	2.3300	4	40
28	2.7600	3.5000	40
29	2.9800	5.5000	19
30	2.5600	3.5000	40

Figure 4.1 Results of affinity propagation



	IPK	Durasi PI (Tahun)	Centr
1	3	2	
2	3.1200	1.5000	
3	3.0500	2	
4	3.3300	1.5000	
5	3.5400	1	
6	2.9000	3.5000	
7	3.8000	0.5000	
8	3.9000	0.5000	
9	2.1200	4.5000	
10	2.8900	3	
11	3.5900	1	
12	3.2200	2.5000	
13	3.4500	3	
14	2.3300	4	
15	2.7600	3.5000	
16	2.9800	5.5000	
17	2.5600	3.5000	
18	2.6500	4.5000	
19	2.4300	5	
20	3.2500	1	
21	3.3300	1.5000	
22	3.6200	1	
23	3.7700	0.5000	
24	3.5900	1	
25	3.2200	2.5000	
26	3.4500	3	
27	2.3300	4	
28	2.7600	3.5000	
29	2.9800	5.5000	
30	2.5600	3.5000	

Figure 4.2 Results of K-Means

Table 4.1 Cluster testing in Affinity Propagation

No	Data		Exemplar				
	GPA	Duration (sms)	1st Test	2nd Test	3th Test	4th Test	5th Test
1	3	2	3	3	3	3	3
2	3,12	1,5	3	3	3	3	3
3	3,05	2	3	3	3	3	3
4	3,33	1,5	3	3	3	3	3
5	3,54	1	22	22	22	22	22
6	2,9	3,5	40	40	40	40	40
7	3,8	0,5	22	22	22	22	22
8	3,9	0,5	22	22	22	22	22
9	2,12	4,5	19	19	19	19	19
10	2,89	3	40	40	40	40	40
11	3,59	1	22	22	22	22	22
12	3,22	2,5	3	3	3	3	3
13	3,45	3	40	40	40	40	40
14	2,33	4	40	40	40	40	40
15	2,76	3,5	40	40	40	40	40
16	2,98	5,5	19	19	19	19	19
17	2,56	3,5	40	40	40	40	40
18	2,65	4,5	19	19	19	19	19
19	2,43	5	19	19	19	19	19
20	3,25	1	22	22	22	22	22
21	3,33	1,5	3	3	3	3	3
22	3,62	1	22	22	22	22	22
23	3,77	0,5	22	22	22	22	22
24	3,59	1	22	22	22	22	22
25	3,22	2,5	3	3	3	3	3
26	3,45	3	40	40	40	40	40
27	2,33	4	40	40	40	40	40
28	2,76	3,5	40	40	40	40	40
29	2,98	5,5	19	19	19	19	19
30	2,56	3,5	40	40	40	40	40
31	3,76	1	22	22	22	22	22
32	3,9	0,5	22	22	22	22	22
33	3,43	2	3	3	3	3	3
34	3,89	1	22	22	22	22	22
35	3,59	2	3	3	3	3	3
36	3,22	1	22	22	22	22	22
37	3,45	1	22	22	22	22	22
38	2,33	1,5	3	3	3	3	3
39	2,76	1	22	22	22	22	22
40	2,98	3,5	40	40	40	40	40
41	2,56	1,5	3	3	3	3	3
42	3,76	1	22	22	22	22	22
43	3,9	0,5	22	22	22	22	22
44	3,43	1	22	22	22	22	22
45	3,89	0,5	22	22	22	22	22
46	3,22	2,5	3	3	3	3	3
47	3,45	3	40	40	40	40	40
48	2,33	4,5	19	19	19	19	19
49	2,76	3,5	40	40	40	40	40
50	2,98	5,5	19	19	19	19	19

Table 4.2 Cluster testing K-Means:

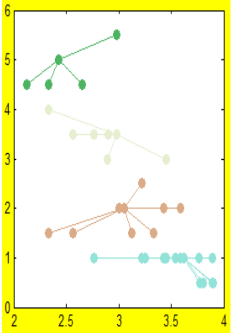
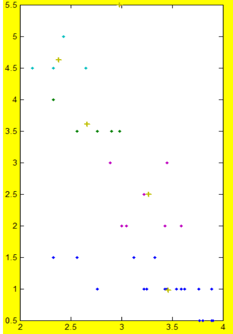
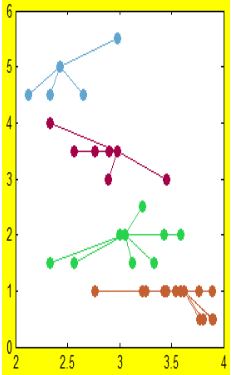
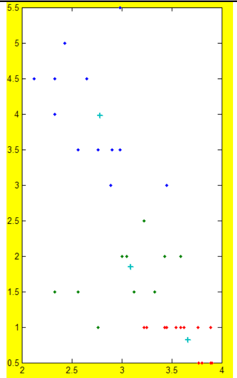
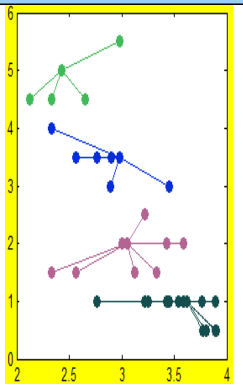
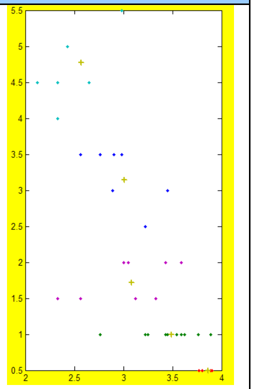
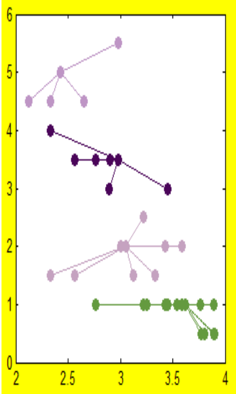
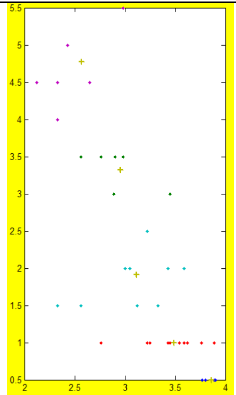
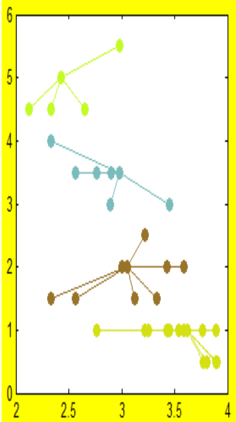
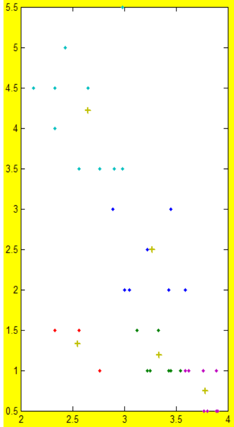
No	Data		Centroid				
	GPA	Duration (sms)	1st Test	2nd Test	3th Test	4th Test	5th Test
1	3	2	5	5	5	2	4
2	3,12	1,5	1	5	3	5	4
3	3,05	2	5	5	5	2	4
4	3,33	1,5	1	5	3	5	4
5	3,54	1	2	3	2	4	1
6	2,9	3,5	3	4	1	1	2
7	3,8	0,5	2	3	4	4	1
8	3,9	0,5	2	3	4	4	1
9	2,12	4,5	4	4	1	1	3
10	2,89	3	3	2	5	2	2
11	3,59	1	2	3	2	4	1
12	3,22	2,5	5	2	5	2	5
13	3,45	3	5	2	5	2	5
14	2,33	4	3	4	1	1	2
15	2,76	3,5	3	4	1	1	2
16	2,98	5,5	4	1	1	1	3
17	2,56	3,5	3	4	1	1	2
18	2,65	4,5	4	4	1	1	3
19	2,43	5	4	1	1	1	3
20	3,25	1	1	3	2	5	1
21	3,33	1,5	1	5	3	5	4
22	3,62	1	2	3	2	4	1
23	3,77	0,5	2	3	4	4	1
24	3,59	1	2	3	2	4	1
25	3,22	2,5	5	2	5	2	5
26	3,45	3	5	2	5	2	5
27	2,33	4	3	4	1	1	2
28	2,76	3,5	3	4	1	1	2
29	2,98	5,5	4	1	1	1	3
30	2,56	3,5	3	4	1	1	2
31	3,76	1	2	3	2	4	1
32	3,9	0,5	2	3	4	4	1
33	3,43	2	5	2	5	2	5
34	3,89	1	2	3	2	4	1
35	3,59	2	5	2	5	2	5
36	3,22	1	1	3	2	5	1
37	3,45	1	2	3	2	4	1
38	2,33	1,5	1	5	3	5	4
39	2,76	1	1	5	3	5	4
40	2,98	3,5	3	4	1	1	2
41	2,56	1,5	1	5	3	5	4

No	Data		Centroid				
	GPA	Duration (sms)	1st Test	2nd Test	3th Test	4th Test	5th Test
42	3,76	1	2	3	2	4	1
43	3,9	0,5	2	3	4	4	1
44	3,43	1	2	3	2	4	1
45	3,89	0,5	2	3	4	4	1
46	3,22	2,5	5	2	5	2	5
47	3,45	3	5	2	5	2	5
48	2,33	4,5	4	4	1	1	3
49	2,76	3,5	3	4	1	1	2
50	2,98	5,5	4	1	1	1	3

4.2 Results

Based on clustering test which has been done on both algorithm, then obtained the results as shown in the table below:

Table 4.3 Result of both algorithm clustering

Tes t	Algorithm	
	Affinity propagation	K-Means
1		
2		
3		
4		
5		

From the results of existing clusters in the table, it can be seen some results from both of algorithm, they are:

a) In affinity propagation algorithm, obtained the members and the number of clusters that remains the same after 5 times of testing, they are 4 clusters.

b) In K-Means algorithm, there are members and the number of clusters which is different after 5 times of testing

c) In affinity propagation formed 4 data clusters based on Duration of Bachelor-Thesis completion variable, where first cluster revolves around 0 to 1 semester, second cluster ranged from 1 to more than 2 semesters, the third cluster is at 3 to 4 semesters, and fourth cluster ranged from 4 to more than 5 semesters

d) In K-Means, formed a variety of clustering results. Can be seen in the table of K-Means testing before, where there are centroid values which are different in each test performed.

e) Apparently, there is a relationship between two variables which in used, it is between GPA student and Duration of Bachelor-Thesis completion. From the results contained in tables, it can be seen that for students who have GPA above 3 to 4 have a tendency to do Bachelor-Thesis faster than other, which is within less from 1 until 2 semesters. While for students who have a GPA less than 3 have a longer time to solve Bachelor-Thesis, which is within a period of 2 until more than 4 semesters.

f) From the latest test, it can also be seen that affinity propagation can provide a static clusters. Where in affinity propagation does not occur the changing of value of exemplar, thus producing the same exact number of clusters at each time of testing

g) From the latest test it also can be seen, that K-Means gives results of cluster which tend to be unstable, where the number of clustering in each trial subject changes. So, the value of centroid also can’t be similar at each time of testing.

h) By comparing the two algorithms, can be known that affinity propagation could provide more optimal clusters than K-Mean algorithm.

5. CONCLUSION

Affinity propagation can be regarded as a cluster method which implements the message-passing algorithm, where data clusters are formed based on the delivered message and the received massage between each data point. After the testing stage is done, it can be conclude that:

a) Affinity propagation gives the result of data cluster more accurate and effective than K-Means, it can be seen from the testing table which showing that the value of affinity propagation exemplar has not changed at all after five trials. While K-Means, gives values of its centroid are different after five trials.

b) There is a relationship between GPA and Duration of Bachelor-Thesis completion in Gunadarma University students, it can be seen from the results of data clustering, which is for student who have GPA above 3 to 4 have a tendency to finish their Bachelor-Thesis faster, which is less than 1 until 2 semesters. While other students who have GPA less than 3 have a longer time to finish their Bachelor-Thesis, within a period of 2 until more than 4 semesters.

There are some drawbacks caused by the limited knowledge possessed by the authors in making the program, such as a simple interface and the workings of programs. Testing could be more optimized when using more data in testing, so that the clusters formed will be more clearly. Therefore, it is expected the suggestions and ideas which can be extended to improve the implementation and development of this algorithm to become better and useful.

REFERENCES

[1] Agusta, Yudi. (2007). *K-Means Penenrapan, Permasalahan, dan Metode Terkait.* Jurnal Sistem dan Informatika.

[2] Frey, Brendan J & Delbert Dueck. (2004). *Mixture Modelling by Affinity Propagation*. University of Toronto.

[3] Knight, Andrew. (2000). *Basics of MATLAB and Beyond*. New York: Chapman & Hall CRC.

[4] Leone, Michele.(2007). Clustering by Soft-Canstraint Affinity Propagation: Applications to Gene Expression Data. Institute for Scientific Interchange.

[5] Otto, S.R & J.P Denier. (2005) An Introduction to Programming and Numerical Methods in MATLAB. London: Springer-Verlag.

[6] Redmond, Patrick. (2007). Affinity Propagation and Other Data Clustering Techniques.

[7] Register, Andy H. (2007). *A Guide to MATLAB Object Oriented Programming*. Georgia: Schitech Publishing, Inc.

[8] Thavikulwat, Precha. (2008). Affinity Propagation: A Clustering Algorithm for Computer- Assisted Business Simulations and Eperiential Exercises. Towson University.

[9] Wilson, Howard B. (2003). *Advanced Mathematics and Mechanic Using MATLAB*. New York: Chapman & Hall CRC.

[10] "Algoritma". http://www.wikipedia.com. (accessed on June- 2- 2011)

[11] "Optimasi Titik Pusat K-Means Dengan Algoritma Genetika". http://www.docs.google. com. (accessed on July-15- 2011)

AUTHOR PROFILES:

Achmad Benny Mutiara was born in Jakarta, in 1967, and is a Professor of computer science at Gunadarma University. He received the B.S degree in Dept. of Physics from University of Indonesia and Dept. of Informatics Engineering from Gunadarma University, Indonesia, in 1991. He also received the M.S and PhD degrees in Computation from Universitaet Goettingen, Germany, in 1996 and 2000, respectively. He is Dean, Faculty of Computer Science and Information Technology at Gunadarma University. His current interests are Computer Modeling and Simulation (esp. Molecular Dynamics Simulation and Monte Carlo), parallel computing (PC-Clustering), and Computational Science.

Rina Refianti is with Faculty of Computer Science and Information Technology at Gunadarma University. She is a Assistant Professor, received the B.S. and M.S. degree in Dept of Information system from Gunadarma University, in 1991 and 2003, respectively.

A. Juarna is a Assoc.Prof and combinatorlist at Faculty of Computer Science and Information Technology, Gunadarma University, Indonesia. He got his Ph.D *dual degree* in Combinatorics from Universite de Bourgogne-France under supervising of Prof. Vincent Vajnovszki and from Gunadarma University under supervising of Prof. Belawati Widjaja. Some of his papers were presented in some conference such as Words-2005, CANT-2006, GASCom-2006, and some others are published in some journals or research reports such as CDMTCS-242 (2004), CDMTCS-276 (2006), The Computer Journal 60(5)-2007, Taru-DMSC 11(2)-2008.