IT FOR SOCIETY, Vol. 04, No. 01
ISSN 2503-2224

IMPLEMENTATION OF K-MEANS ALGORITHM FOR INFORMATION TECHNOLOGY FRESHMAN CLASS DIVISION

Arfan As'Sidiq[1]
Rila Mandala[2]
Department of Computer Science
President University,
Cikarang, Bekasi, 17550, Indonesia
[1]Arfan.assidiq@gmail.com
[2]rilamandala@president.ac.id

Abstract

Almost all universities divide their IT freshman into classes randomly or based on students score, either their score during the selection test held by the university or National Examination score. Universities often find case that a class consists of all 'smart' students and a class consists of all 'lazy' students. This thesis intends to create an application to help universities divides their Information Technology freshman into classes based on freshman competency and experience about Information Technology (IT) on the senior high school. The experiment is conducted by collecting data IT students who are not in the first semester. The data consists of their experience about IT as well as other knowledge fields and their current GPA. The results of the experiment show that from 50 data samples collected, the application correctly predicts 34 students GPA range based on respondents competency with IT and other knowledge fields during their study in senior high school.

1. Introduction

The academic score students get at high school is not the representation of student's knowledge about IT competency for reasons. The scores students get may not purely the student's work since the students may cheat while doing the test. Besides, an academic test is more likely to test the students memory of what the students have studied instead of the students capability to elaborate the problems. Instead of using academic score, classifying the students by the students competency and experience about IT may give better consideration to divide the IT freshman into classes.

This thesis will develop an application to do clustering IT freshman competency using K-Means algorithm. The data sample is collected by surveying IT students who are not in the first semester to compare the results of the application accuracy classifying IT freshman in real world.

The problems that will be solved are: 1. How to get information about university's IT freshman competency, 2. How to divide the IT freshman into classes fot the first semester based on the freshman competency.

These can be solved by having an application to clusters the freshman based on the freshman knowledge and experience about IT during their study at high school.

The objectives of this thesis are:

- Collect the IT competency data from the IT freshman.
- Get information about freshman competency about IT by processing data that have collected.

- Give suggestion for university how to divide the IT feshman into classes for the first semester.

2. Limitations

This application consists of two parts: Web Survey and Python Program
The Web Survey provides survey form for IT freshman to collect the data about freshman competency that is developed using Django Framework. The questions for the survey are divided into two categories: Related to IT competency and Not related to IT competency.
The answer of the question is a multiple choice: Yes or No.
The data that have been collected can be downloaded from Web Survey as CSV file.
The clustering results can be retrieved by processing the CSV file by running the Python Program.
The accuracy of the clustering depends on how accurate the questions given to the freshman about knowledge and experience about IT.

3. Methods

This section explains the methods used in the development of the application.

3.1 Clustering

Clustering is the grouping of a particular set of objects based on their characteristics and aggregating according to their similarities. Regarding to data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis.

This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships [1].

3.2 K-Means Algorithm

K-Means algorithm follows a simple way to classify given data set by associate it to the nearest centroid as its group based on each data attributes, then classify those data attributes into a set of predetermined K-clusters.
Generally, the K-Means algorithm works by taking a unique attribute from each data given as a data point, then assign these data points to the nearest cluster centroid, after that the centroids will begin moving to better fit the clusters themselves. These data points and centroids will keep moving until the algorithm find the minimum fitting error between the data sets and centroids by updating the centroids to be the mean value of the clusters repeatedly in each loop until no movement needed anymore [2]. Figure 1 below shows K-Means algorithm flowchart.

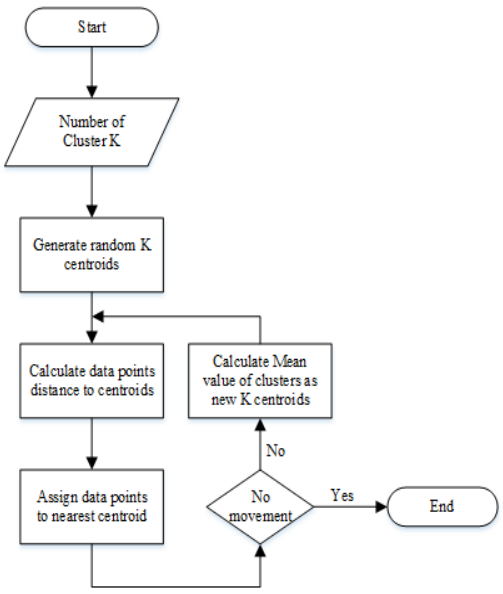

Figure 1 Basic K-Means Algorithm Flowchart

3.3 Python Programming Language

Python is a widely used high-level, general-purpose, interpreted, and dynamic programming language. Created by Guido von Rossum in 1980, its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than possible in the other programming language such as C++ or Java. The language provides constructs intended to enable writing clear programs on both small and large scale.

Python interpreters are available for many operating systems, allowing Python code to run on a wide variety of systems. Python is based on C language or usually called CPython. Although there are many programming languages that are combined with Python, there are mainly two major Python implementations: Jython or Java-Based Python and IronPython that is adapted in .NET framework of Microsoft[3].

3.4 Django Framework

Django is a free and open-source web framework, written in Python, which follows the model-view-template (MVT) architectural pattern. It is maintained by the Django Software Foundation (DSF), an independent organization established as a 501(c)(3) non-profit.

Django's primary goal is to ease the creation of complex, database-driven websites. Django emphasizes reusability and "pluggability" of components, less code, low coupling, rapid development, and the principle ‘don't repeat’[4]. Python is used throughout, even for settings files and data models. Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models [4].

4. Experimental Results

Tests are done in order to evaluate the effectiveness of the proposed methods in previous section. The data samples are taken from IT students who are not in the first semester. The x coordinates represents the number of respondent’s points of IT competency and the y coordinates represents the number of points the respondent’ss points of non IT competency. The survey consists of 27 questions: 13 IT related competency questions and 14 non-IT related competency questions. Every ‘Yes’ answer of each question gives 1 point to respondent competency point. Centroid of C1 is located on both maximum points multiplied by ¾ (9.75 , 10.5). Centroid of C2 is located on maximum point of IT competency multiplied by ¾ and maximum point of non IT competency multiplied by ¼ (9.75 , 3.5). Centroid of C3 is located on maximum point of IT related competency multiplied by ¼ and maximum point of non IT related competency multiplied by ¾ (3.25 , 10.5). Centroid of C4 is located on both maximum points multiplied by ¼ (3.25 , 3.5). The iteation of clustering is done in fifth iteration. Test results are described in tables below:

Table 1 Centroids Coordinate

Iteration	Centroids	Coordinate	
		x	y
I	C1	9.75	10.5
	C2	9.75	3.5
	C3	3.25	10.5
	C4	3.25	3.5
II	C1	11.0	11.0
	C2	5.5	5.5
	C3	1.71	6.47
	C4	1.03	3.63
III	C1	11.0	11.0
	C2	4.6	5.2
	C3	1.63	6.21
	C4	0.68	3.44
IV	C1	11.0	11.0
	C2	4.6	5.2
	C3	1.63	6.21
	C4	0.68	3.44
V	C1	11.0	11.0
	C2	4..14	5.14
	C3	1.22	6.0
	C4	0.74	2.95

Table 2 K-Means Clustering Results

No	Name	Coordinate		Distance from				Cluster Member
		x	y	C1	C2	C3	C4	
1	Haikal	2	6	10.3	2.31	0.62	3.09	C3
2	Fira	0	6	12.08	4.23	1.22	3.14	C3
3	Grace	1	6	11.18	3.26	0.22	3.06	C3
4	Neo	0	9	11.18	5.66	3.24	6.1	C3
5	Kristoforus	6	6	7.07	2.05	4.78	6.08	C2
6	Fransiscus	0	6	12.08	4.23	1.22	3.14	C3
7	Leonardo	3	7	8.94	2.18	2.04	4.64	C3
8	Melissa	4	4	9.9	1.15	3.42	3.42	C2
9	Enda	5	5	8.49	0.87	3.91	4.73	C2
10	Rai	2	3	12.04	3.03	3.1	1.26	C4
11	Rafqi	2	3	12.04	3.03	3.1	1.26	C4
12	Lia	2	6	10.3	2.31	0.78	3.3	C3
13	Aldo	0	6	12.08	4.23	1.22	3.14	C3
14	Aulia	1	6	11.18	3.26	0.22	3.06	C3
15	Desriel	3	5	10.0	1.15	2.04	3.05	C2
16	NadyaW	2	5	10.82	2.14	1.27	2.41	C3
17	Raihan	0	0	15.56	6.6	6.12	3.04	C4
18	Vera	4	7	8.06	1.87	2.95	5.2	C2
19	Daffa	1	5	11.66	3.14	1.02	2.07	C3
20	Tedy	0	5	12.53	4.14	1.58	2.18	C3
21	Muhammad	2	8	9.49	3.57	2.15	5.2	C3
22	Kelvin	0	4	13.04	4.29	2.34	1.28	C4
23	Nikita	2	7	9.85	2.84	1.27	4.24	C3
24	Nadya	0	3	13.6	4.66	3.24	0.74	C4
25	Zah	2	3	12.04	3.03	3.1	1.26	C4
26	Fariz	0	4	13.04	4.29	2.34	1.28	C4
27	Hawarti	1	3	12.81	3.8	3.01	0.26	C4
28	MuhArfan	0	5	12.53	4.14	1.58	2.18	C3
29	Abdul	0	4	13.04	4.29	2.34	1.28	C4
30	Rismayanti	0	5	12.53	4.14	1.58	2.18	C3
31	Yayang	1	0	14.87	6.02	6.0	2.96	C4
32	Arlan	2	4	11.4	2.42	2.15	1.64	C4
33	Muhandika	2	7	9.85	2.84	1.27	4.24	C3
34	Gesang	3	5	10.0	1.15	2.04	3.05	C2
35	Alfian	4	4	9.9	1.15	3.42	3.42	C2
36	Zulhan	2	4	11.4	2.42	2.15	1.64	C4
37	Labib	11	11	0.0	9.02	10.98	13.04	C1
38	Novita	0	4	13.04	4.29	2.34	1.28	C4
39	Aziz	2	5	10.82	2.14	1.27	2.41	C3
40	MuhammadMulya	1	4	12.21	3.34	2.01	1.08	C4
41	Aftiani	0	3	13.6	4.66	3.24	0.74	C4
42	Raka	0	0	15.56	6.6	6.12	3.04	C4
43	Afifah	1	5	11.66	3.14	1.02	2.07	C3
44	Umar	2	7	9.85	2.84	1.27	4.24	C3
45	NurWidianti	1	5	11.66	3.14	1.02	2.07	C3
46	Amelia	0	4	13.04	4.29	2.34	1.28	C4
47	Zamzam	2	6	10.3	2.31	0.78	3.3	C3
48	Tomikusumo	1	4	12.21	3.34	2.01	1.08	C4
49	Susi	0	2	14.21	5.2	4.18	1.2	C4
50	Aditya	2	5	10.82	2.14	1.27	2.41	C3

Table 3 Class Dividing Results

Classes	Name	Cluster
Class 1	Labib	C1
	Kristoforus	C2
	Melissa	C2
	Enda	C2
	Alfian	C2
	Haikal	C3
	Fira	C3
	Grace	C3
	Neo	C3
	Fransiscus	C3
	Leonardo	C3
	Lia	C3
	Aldo	C3
	Aulia	C3
	NadyaW	C3
	Daffa	C3
	Aditya	C3
	Rai	C4
	Rafqi	C4
	Raihan	C4
	Kelvin	C4
	Nadya	C4
	Zah	C4
	Fariz	C4
	Hawarti	C4
	Abdul	C4
Class 2	Desriel	C2
	Vera	C2
	Gesang	C2
	Tedy	C3
	Muhammad	C3
	Nikita	C3
	MuhArfan	C3
	Rismayanti	C3
	Muhandika	C3
	Aziz	C3
	Umar	C3
	NurWidianti	C3
	Zamzam	C3
	Yayang	C4
	Arlan	C4

Table 1 GPA Categories

The predictions of the clustering evaluation are as follows:

- Freshman who are included in C1 are having category A+ or category A GPA.
- Freshman who are included in C2 are having category A+ or category A GPA.
- Freshman who are included in C3 are having category A or category B GPA
- Freshman who are included in C4 are having category B or C GPA.

	Raka	A	Having category B or C GPA	Not as expected
	Amelia	A	Having category B or C GPA	Not as expected
	Tomi	B	Having category B or C GPA	As expected
	Susi	B	Having category B or C GPA	As expected

The comparison between the prediction and the data sample is shown in table below:

Table 2 Evaluation Results

Cluster Member	Name	GPA Categories	Expected Results	Results
C1	Labib	A	Having category A+ or A GPA	As expected
C2	Kristoforus	B	Having category A+ or A GPA	Not as expected
	Melissa	B	Having category A+ or A GPA	Not as expected
	Enda	B	Having category A+ or A GPA	Not as expected
	Desriel	B	Having category A+ or A GPA	Not as expected
	Vera	B	Having category A+ or A GPA	Not as expected
	Gesang	C	Having category A+ or A GPA	Not as expected
	Alfian	B	Having category A+ or A GPA	Not as expected
C3	Haikal	B	Having category A or B GPA	As expected
	Fira	B	Having category A or B GPA	As expected
	Grace	B	Having category A or B GPA	As expected
	Neo	B	Having category A or B GPA	As expected
	Fransiscus	B	Having category A or B GPA	As expected
	Leonardo	B	Having category A or B GPA	As expected
	Lia	B	Having category A or B GPA	As expected
	Aldo	B	Having category A or B GPA	As expected
	Aulia	B	Having category A or B GPA	As expected
	NadyaW	A	Having category A or B GPA	As expected
	Daffa	B	Having category A or B GPA	As expected
	Tedy	B	Having category A or B GPA	As expected
	Muhammad	B	Having category A or B GPA	As expected
	Nikita	B	Having category A or B GPA	As expected
	MuhArfan	A	Having category A or B GPA	As expected
	Rismayanti	B	Having category A or B GPA	As expected
	Muhandika	A	Having category A or B GPA	As expected
	Aziz	B	Having category A or B GPA	As expected
	Afifah	A	Having category A or B GPA	As expected
	Umar	B	Having category A or B GPA	As expected
	NurWidianti	B	Having category A or B GPA	As expected
	Zamzam	C	Having category A or B GPA	Not as expected
	Aditya	A	Having category A or B GPA	As expected
C4	Rai	B	Having category B or C GPA	As expected
	Rafqi	B	Having category B or C GPA	As expected
	Raihan	C	Having category B or C GPA	As expected
	Kelvin	B	Having category B or C GPA	As expected
	Nadya	B	Having category B or C GPA	As expected
	Zah	A	Having category B or C GPA	Not as expected
	Fariz	A	Having category B or C GPA	Not as expected
	Hawarti	A	Having category B or C GPA	Not as expected
	Abdul	B	Having category B or C GPA	As expected
	Yayang	A+	Having category B or C GPA	Not as expected
	Arlan	C	Having category B or C GPA	As expected
	Zulhan	C	Having category B or C GPA	As expected
	Novita	A+	Having category B or C GPA	Not as expected
	MuhammadMulya	B	Having category B or C GPA	As expected
	Aftian	A	Having category B or C GPA	Not as expected

Based on the results, C1 only have one member and the results is as expected. The results of all 7 members on C2 are not as expected. C3 have 23 members and only 1 member is not as expected. C4 have 19 members and have 11 members that the results is as expected. From 50 real data samples, there are 34 data sample as predicted by the K-Means algorithm. The python program has 68% of accuracy(34 / 50 x 100) based on 50 data samples to predict freshman GPA range the freshman will get based on their experience with IT during the freshman study in senior high school.

5. Discussion

This section presents discussion on why the clustering is using K-Means algorithm, Django web framework, and Python programming language.

- K-Means algorithm

This thesis using K-Means Algorithm utilizes clustering method for classifying the IT freshman rather than another method such as Fuzzy C-Means.

The advantage of K-Means algorithm:
System can maximize the classification of the IT freshman into a single category membership rather than instead of classifies the IT freshman into two or more categories with its certain degree of membership like using Fuzzy C-Means.

- Python Programming Language

The clustering algorithm is applied using Python programming language for the reason of tidier and simpler syntax, faster execution, and more intuitive compare to other programing language.

The advantage of using Python is:
System can be developed faster with more understandable code, fast

processing, and easier for debugging purpose during the development.

- Django Web Framework

The Web Survey is developed using Django as its framework for the reason of faster performance, lighter, and there is a built-in admin dashboard.

The advantage of using Django is:
Since Django is a web framework that is also written in Python programing language, Django inherits all the advantages of Python programing language. The Django framework is also more customable both its interface and the back-end processing compare to other framework.

6. Conclusions

First, this thesis develop an application that consists of two parts : Web Survey and Python Program to help university classify IT freshman competency based on their knowledge and their experience about information technology before join university as well as their knowledge and competency in other fields for dividing them into classes on the first semester.
Second, the number of cluster is fixed to four clusters: C1 (freshman with high IT competency and high non-related IT competency), C2 (freshman with high IT competency), C3 (freshman with high competency in their fields), and C4 (freshman with average competency).
Third, the K-Means algorithm has 68% of accuracy based on the 50 data samples in order to predict freshman GPA range the freshman will get based on their experience with IT during the freshman study at senior high school
In the future, the Web Survey will have better user interface and better user experience for the user and the K-Means algorithm become more flexible by automatically assigning the number of centroid and its coordinate rather than predifined centroids.

7. Acknowledgements

The author would like to thank Mr. Rila Mandala, Ph.D. as the advisor of this thesis, Mr. Nur Hadisukmana, M.Sc., as the Program Head of Information Technology, and Mr. Rikip Ginanjar, M.Sc.

References

[1] *Tan, P., Steinbach, M., & Vipin, K*. 2016. Introduction to Data Mining 1st edition. India: Pearson.

[2] *Han, J., & Kamber, M*. 2006. Data Mining: Concepts and Techniques 2nd edition. San Fransisco: Elsevier.

[3] Wikipedia. *Python Programming Language.* Accessed October 13, 2018, from Wikipedia: https://en.wikipedia.org/wiki/Python_(programming_language).

[4] Wikipedia. *Django (web framework).* Accessed October 13, 2018, from Wikipedia: https://en.wikipedia.org/wiki/Django_(web_framework).