Implementation of K-Means Clustering Method to Distribution of High School Teachers

Triyanna Widiyaningtyas, Martin Indra Wisnu Prabowo, M. Ardhika Mulya Pratama
Electrical Engineering Departement
Universitas Negeri Malang
Malang, Indonesia
triyannaw.ft@um.ac.id, martinindra19@gmail.com, hyoudou.dhika@gmail.com

***Abstract*—Currently, the government is still having difficulties in distributing teachers. The current problem is not just about less teachers, but also more teachers in some cities. The problem of unequal distribution of teachers then became dependent on local government. The distribution of teachers now can not be centralized because of the decentralization system implemented in Indonesia. Clustering in data mining is useful for finding distribution patterns within a dataset that is useful for data analysis processes. Using clustering, identifiable densely populated areas, overall distribution patterns and attractive associations between data attributes. The purpose of this research is to apply *k*-means clustering algorithm to analyze distribution of high school teachers in Indonesia. This research uses three steps, namely dataset selection, preprocessing data, and application of k-means clustering. Testing is done by using *k* cluster, that is *k* = 12. The cluster results are analyzed to classify clusters into 3 categories, namely less, enough, and more teachers. Testing results obtained data Sum of Squared Error (SSE) with percentage 87.15%. While the clustering results produce clusters 3 and 5 in the category of less teachers. Cluster 1 and 9 in the category of enough teachers. While cluster 2,4,6,7,8,10,11,12 in the category of more teachers. Based on the results obtained it can be concluded that the accuracy of the algorithm used with 12 clusters is very high. The results of this clustering analysis can also be used as a reference for the distribution of teachers to region with less teachers, so as to solve the issue of uneven distribution of teachers.**

Keywords— k-means clustering; student ratio; teacher ratio

I. INTRODUCTION

The Ministry of Education and Culture is responsible for the implementation, coaching, and management of educational and cultural activities in the Republic of Indonesia. Science and technology development that happens in this world, not escape from the existing educational role. Education is something that is urgently needed to educate the next generation. Not only that education is also a very viral thing for people who are serious to deepen knowledge and sharpen.

According to Dewey, education is the learning of knowledge, skills, and habits of a group of people who are passed down from one generation to the next through teaching, training, or research. Education often occurs under the guidance of others, but also enables self-taught [1]. Education also can not be separated from the role of teachers and learners. In Indonesia, the number of learners is very much, even in some areas have a lot of learners but lack of teachers. The spread of teachers is an important element in the quality control of education. According to Goverment Regulation Number 74/2008 on Teachers, the ratio of teachers to students is 1 in 20. A teacher is considered effective to teach 20 students in elementary, junior high and high school.

Based on data from The State Employee Agency (SEA), DKI Jakarta is the province with the highest number of civil teachers, North Borneo is the province with the fewest teachers. SEA data on January 24, 2017 mentioned the distribution of teacher composition at provincial level at most in DKI Jakarta Province with 33.037 teachers. For the lowest provincial teacher distribution composition is located in the area of North Borneo Province with a total of 1.116 teachers. Furthermore at the district level, the distribution of most teachers are in Bandung District with the number of 11.657 teachers. On the contrary, the lowest number of teachers is in the Arfak Mountains region which is included in the scope of West Papua Province, with 42 teachers. Data on teacher distribution at provincial and district levels shows that teacher distribution has not been equitable.

Another example as written in the web www.cnnindonesia.com in June 2015, research results say in Yogyakarta has over 200 math teachers. This is because the number of math teachers required only 520 people, but the existing math teacher about 726 people [2]. In that case, we can know that for some or even almost all regions in Indonesia, there is an imbalance in the distribution of teachers.

Clustering is an essential task in data mining process which is used for the purpose to make groups or clusters of the given data set based on the similarity between them. K-means is one of clustering algorithm. K-means was first discovered by Lloyd in 1957. Then Forgey in 1965 also published the same method, so-called Lloyd-Forgey. K-means can help solve problems in grouping, such as biological, business, multimedia, and information [3]. The purpose of this algorithm is to divide the data into groups. The received input is the data or object and k of the desired cluster. This algorithm will group the data or objects into the k kinds of the group.

K-Means is a method of data mining that performs the modeling process without supervise and also one of the

978-1-5386-0549-3/17/$31.00 ©2017 IEEE

methods using the method of grouping data in partition. The K-Means method seeks to group the existing data into groups. Where the data contained in a group have the same characteristics with each other but have different characteristics with the data in the other group. Therefore, this method can be used to bridge variations between existing data in a cluster and additionally with data that exist in other clusters. The *k*-means method is more efficient for large data processing because the *k*-means complexity calculation formula is *O (n.k.t)* where *n* is the total number of objects and *k* is the number of clusters and t is the number of iterations. In general, the $k << n$ and $t << n$ values often result in local optimal [4].

Some research related to the clustering method, including Johan Oscar Ongg's research in his journal entitled "Implementation of K-Means Clustering Algorithm for Determining Marketing Strategy of President University". Oscar's research results prove that the application of *k*-means algorithm is very effective in determining the marketing strategy [5]. The clustering algorithm used can map two marketing strategies, that is (1) do a promotion by sending the right marketing team and (2) do a promotion to cities in Indonesia based on the academic ability level of the prospective student.

Another study conducted by Narwati on grouping students using K-means algorithm. This study classifies students based on academic data using clustering techniques, so it is expected to provide information for those concerned. This clustering technique algorithm begins with random selection of K, which is the number of clusters that want to be formed from the data that will be in cluster. That are the admission test score of the student and Grade Point Average (GPA). The system is made to display the cluster of student academic data. Namely the pattern of student achievement that cluster fixed, down and up, and can be seen from the cource of study, hometown and high school. The result of the case study can be obtained the information of students who remain in clusters such as the initial entry as much as 422 (45, 085%), students who climbed clusters as much as 284 (30.342%) and students who fell clusters as much as 230 (24.573%) [6].

Another research of *k*-means clustering is presented by Kaur. There are certain limitations in *k*-means clustering algorithm such as it takes more time for execution. So in order to reduce the execution time, they are using the Ranking Method. And also shown that how clustering is performed in less execution time as compared to the traditional method. The proposed work represents ranking based method that improved *k*-means clustering algorithm performance and accuracy. The experimental results demonstrated that the proposed ranking based *k*-means algorithm produces better results than that of the existing *k*-means algorithm.

Based on this problem, this study aims to apply the *k*-means clustering algorithm in conducting the distribution of high school teachers in Indonesia, so as to provide a solution of equal distribution of teachers in accordance with the clusters formed. K-means clustering is one method in data mining that can be used to overcome this problem. K-means clustering is an algorithm that aims to divide data into several groups. In this method, the entered data will be grouped into a certain class, with the data having no specific label or class. This algorithm will group data or objects into several groups. K-means clustering method can classify teacher equalization data into three categories, namely less, enough, and more teachers.

II. METHODE

This study uses three stages: (1) dataset selection, (2) preproccessing data, and (3) *k*-means clustering, which is illustrated in Fig. 1 below.

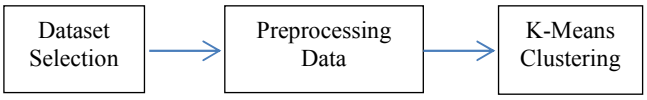

Fig. 1. Three stage of study

A. Dataset Selection

The data used in this study is the data on the number of students and the number of teachers, namely the data of equalization of students entitled "Ratio of School Students" and the distribution of teachers in Indonesia entitled "Ratio of School Teachers". These data are sourced from the website www.data.go.id, which are ratio of school student and ratio of school teacher. Both data have a ratio of academic year 2014/2015 and 2015/2016 to every province and district/city across Indonesia with data totaling 549 samples and 8 attributes.

Table 1 and Table 2 below illustrate the dataset from ratio of school student and ratio of school teacher that used as the initial dataset of the study.

TABLE 1. THE DATASET OF SCHOLL STUDENT RATIO

Region	Latitude	Longitude	Elemantary School (SD)	Junior High School (SMP)	High School (SMA)	Vocational High School (SMK)	Special Education (PLB)
Indonesia	-2.21962	117.5647	175.65	272.84	349.29	350.95	57.74
Aceh	3.93002	97.8333	141.95	191.13	280.49	248.55	46.87
Bali	-8.40952	115.1889	166.57	494.03	502.27	527.86	114.25
Bangka Belitung	-2.74105	106.4406	200.16	289.15	346.94	409.57	87.89
Banten	-6.40582	106.0640	267.26	320.78	337.44	361.81	57.49
Bengkulu	-3.79285	102.2608	163.27	210.48	355.07	290.37	74.07
D.I. Y	-7.79558	110.3695	158.54	301.25	323.55	376.73	62.64
D.K.I. Jakarta	-6.20876	106.8456	306.99	349.02	350.21	354.62	60.84
Gorontalo	0.54354	123.0568	145.46	154.36	462.98	370.86	114.13
Jambi	-1.58942	103.6091	162.79	196.59	332.36	284.91	102.64

TABLE 2. THE DATASET OF SCHOOL TEACHER RATIO

Region	Latitude	Longi-tude	Elemen tary School (SD)	Junior High School (SMP)	High School (SMA)	Voca-tional High School (SMK)	Special Educa-tion (PLB)
Indonesia	-2.21962	117.5647	11.54	17.78	21.54	19.35	13.03
Aceh	3.93002	97.8333	14.74	21.51	25.85	25.95	12.63
Bali	-8.40952	115.1889	11.06	30.28	29.68	26.52	24.63
Bangka Belitung	-2.74105	106.4406	11.43	15.7	20.00	23.11	17.00
Banten	-6.40582	106.0640	13.55	15.72	16.98	13.58	13.51
Bengkulu	-3.79285	102.2608	11.97	17.12	25.48	21.92	16.67
D.I. Y	-7.79558	110.3695	12.00	21.81	24.06	26.09	17.89
D.K.I. Jakarta	-6.20876	106.8456	15.08	18.03	21.58	15.78	13.65
Gorontalo	0.54354	123.0568	9.62	12.67	23.87	24.08	21.00
Jambi	-1.58942	103.6091	11.70	16.31	21.44	20.11	19.43

Both data are combined and eliminate some attributes. So the data as much as 8784 records and then changed to 5654 records consisting of 514 cities/districts with 514 records and 11 attributes. Then converted to 2056 records consisting of 514 districts/cities with 514 records and 4 attributes.

B. Preprocessing Data

Before doing a cluster on the data, it is necessary preprocessing data first. The preprocessing process includes several steps: (1) data integration, (2) attribute addition, (3) data selection, (4) data conversion, (5) replace missing value, and (6) z-score standarized.

Data Integration

Data integration is a process of combining data from different sources. The unified data aims to support the information required by the user. In this step, 2 dataset integration is done, the data of the ratio of the number of students and the ratio of the number of teachers to schools in cities/districts throughout Indonesia in 2015. Data is integrated with WEKA program, that is by combining dataset 1 and dataset 2, so that data integration result become 549 records and 13 attributes (consisting of attribute of region, latitude, longitude, student_SD, student_SMP, student_SMA, student_SMK, student_PLB, teacher_SD, teacher_SMP, teacher_SMA, teacher_SMK, and teacher_PLB)

Atribut Addition

Attribute additions are done to support and help users to search for information from data with attributes added. The Ministry of Education and Culture Regulation of 2013 which states that each study group has at least 32 to 36 studensts. Based on this regulation, it is necessary to add ideal attributes to the dataset results that have been integrated, so that the 5 ideal attributes will be added to the dataset that has been integrated. The dataset result after added attribute to 18 attributes consisting of region, latitude, longitude, student_SD, student_SMP, student_SMA, student_SMK, student_PLB, teacher_SD, teacher_SMP, teacher_SMA, teacher_SMK, teacher_PLB, ideal_SD, ideal_SMP, ideal _SMA, Ideal _SMK, and ideal_PLB

Data Selection

Selection of data is done to obtain information required by the user. Selection of such data is used to assist the clustering process. From attributes that have been formed in the process of adding attributes need to select the important attributes that are used according to the purpose of clustering.

The initial attribute selection is done by eliminating unused attributes in obtaining the distribution of the teacher, so that the important attributes used are region, teacher_SD, teacher_SMP, teacher_SMA, teacher_SMK, teacher_PLB, and additional attributes containing ideal_SD, ideal_SMP, ideal_SMA, ideal_SMK, and ideal_PLB.

Furthermore, the data is selected again to focus the clustering process, so generate attributes of region, teacher_SMA, teacher_SMK, ideal_SMA, and ideal_SMK. This is done to get maximum cluster results, so that in the clustering process only use data of equal teachers_SMA.

Data Conversion

Data conversion is used to help data adapt to the program. For example by changing the data region name with numbers or numeric. The conversion is done nominal to numeric. Nominal to numeric is done to change the attribute of type nominal to be numeric, that is region attribute. The example of West Aceh district is converted to 1, Southwest Aceh district into 2, and so on. The conversion results are shown in Table 3.

TABLE 3. DATA CONVERSION RESULT

Region	Convert
West Aceh District	1
Southwest Aceh District	2
Aceh Besar District	3
Aceh Jaya District	4
South Aceh District	5

Replace Missing Value

Replace missing values is the process of filling empty or missing data. If the data is of a continuous type, it will be populated by averaging the data in the empty attribute. But if the data is of nominal type, it will be filled with data mode in the attribute.

Replace Missing value is filled with the mean of the attribute. The null type attribute is a PLB attribute with an average of 66.55 and a SMK attribute that has an average of 312.6. Replace missing value results can be seen in Table 4.

TABLE 4. REPLACE MISSING VALUE RESULT

A	B	C	D	E	F
3	140.39	155.49	176.37	256.22	30.50
4	86.11	87.18	172.83	93.86	93.86
5	100.68	184.54	235.03	162.91	66.55
6	160.00	185.97	288.00	176.83	66.55
7	194.11	2214.16	374.95	423.38	205.00
8	103.26	143.17	252.26	443.60	39.38
9	143.04	180.77	270.67	158.64	29.00
10	162.23	192.70	307.62	242.17	56.00
11	166.07	215.12	285.50	312.676	64.50
12	129.44	254.93	414.58	434.85	53.40
13	112.74	116.60	194.45	312.676	30.50
14	188.54	274.37	335.84	312.676	85.00

Z-score Standarized

Z-score is a standard score in the form of a score of a person with a group average. Z-score can usually be a value or in units of deviation standards. Z-score is done if the value of data has a wide range difference.

The Z-score formula can be written as follows:

$$Z = \frac{x - \bar{x}}{s} \quad (1)$$

where $\bar{x}$ = average, s = standard deviation and x = the value to be changed

Z-score is used to change attribute values to be equivalent or have a small range, so making it easier to process. By using Rstudio to help convert the original data into data that has been processed with *Z*-Score. The data has been converted to data with a percentage value. The result of standarized *z*-score conversion can be seen in Table 5.

TABLE 5. THE RESULT OF STANDARIZED Z-SCORE

Region	SMA_ teacher	SMK_ teacher	SMA_ Ideal	SMK_ Ideal
1	1.57312002	-0.21589441	-0.8104286	-0.4672435
2	1.24938047	0.06420588	-0.4855817	-0.7994469
3	2.20021255	0.19796907	-1.7576169	-0.6764139
4	1.58408015	1.09577600	-1.3185906	-1.5802957
5	1.68669133	-0.64378881	-0.2039696	-0.6207700
6	1.96248485	-0.14655311	-0.6354985	-1.0767168
7	0.67300594	-0.03739719	-0.3410092	-0.2710334
8	1.18940942	-1.28111435	-0.8380035	1.1943659
9	1.58751972	0.35464393	-0.7624535	-1.2087707
10	1.18850111	-0.29735032	-0.3249050	-0.4574416
11	1.66435675	-0.93298887	-0.2323437	-0.2270631
12	0.85298710	-0.62625010	-0.3041971	0.2320103

C. *K-means Clustering Algorithm*

The clustering process by using K- means is done after the data has 6 processes in the preprocessing phase of the data. The steps for clustering with the K-means method are [4]:

1. Select the number of clusters *k*
2. Initialize *k* cluster center randomly. Cluster centers are given a random initial value.
3. Allocate all existing object data on the closest cluster. The similarity of two objects can be determined by the distance between the two objects. While the similarity of a data on a cluster is determined by the distance between the data with the center of the cluster. At this stage the calculation of the distance of each data into the center of the cluster. In calculating the distance of all existing data to each center point of the cluster can use Euclidean distance theory with the following formula:

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad (2)$$

where:
$D_{(i,j)}$ = distance of *i*-th data to cluster center *j*
X_{ki} = *i*-th data at attribute *k*-th data
X_{kj} = *j*-th center point at *k*-th attribute

4. Then recalculate the cluster center with the current cluster membership. The cluster center is the average value of all objects/data within the cluster. However, if possible, it can also use the median of the cluster. Thus, the mean is not the only one size that can be used.
5. Repeat the steps of cluster processing before, so there is no change.

In the clustering process with *k*-means, this is done by grouping as many as 12 *k* clusters. The clustering process will result in 12 clusters that will be re-grouped into 3 cluster results to determine the equal distribution of teachers.

III. RESULT AND DISCUSSION

A. *The test with k-means clustering*

The test is done by grouping data using *k*-means method, with the package used is cluster and psych. The test using *k*-means with 514 records and 5 attributes, using cluster $k = 12$. This test get Sum of Squared Error (SSE) data with 87.15% percentage. The result of cluster solution on SSE is shown in Fig. 2, while the result of the plot of *k*-means cluster is shown in Fig. 3.

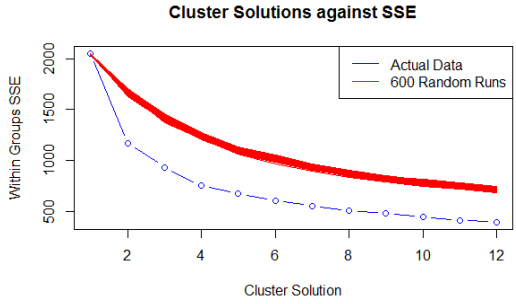

Fig. 2. Cluster solution result to SSE

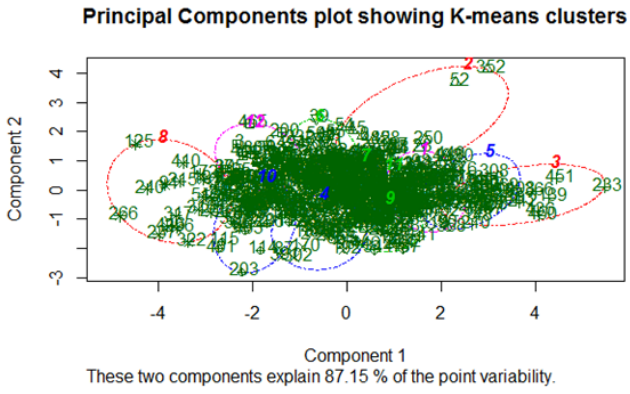

Fig.3. Results of *k*-means cluster plot

Based on the test using *k*-means clustering, it can be seen that 514 records and 5 attributes can produce clustering with cluster number of 12 clusters. The results of the cluster group division can be shown in Table 6 below.

TABLE 6. RESULT OF GROUPING REGION WITH K-MEANS CLUSTERING

Group	Region
1	59,62,71,74,89,97,dst
2	8,19,28,35,40,41,dst
3	138,139,202,210,212,dst
4	43,87,92,96,111,134,dst
5	16,30,90,132,148,dst
6	5,7,10,11,12,17,dst
7	14,31,49,58,70,dst
8	4,82,91,125,155,dst
9	57,61,100,101,105,dst
10	42,63,65,66,81,83,dst
11	18,26,32,34,36,60,dst
12	1,2,3,6,9,13,15,21,dst

B. Cluster Analysis and Cluster Results k=12

The results of the testing with cluster k = 12 with 514 records and 5 attributes can be analyzed by using spreadsheet software. First, the teacher and ideal teacher averages in senior high scholl and vocational high school are calculated. The results are shown in Table 7. Then, the cluster results can be grouped into 3 categories, clusters with less categories, clusters with enough categories, and cluster with more category (if difference > 6). This result is obtained by comparing the average number of teachers and the number of ideal teachers with clustered data. The result of the calculation of the difference in the number of teachers with the ideal number of teachers per cluster is shown in Table 8.

TABLE 7. RESULT OF IDEAL TEACHERS AVERAGE IN EACH CLUSTER

Cluster	AVG (SMA_ teacher)	AVG (SMK_ teacher)	AVG (SMA_ Ideal)	AVG (SMK_ Ideal)
1	13.96870	16.47704	13.90976	10.47503
2	21.83296	17.15481	13.39059	12.92307
3	10.61789	13.25842	17.47887	12.36337
4	15.40658	27.16737	9.451349	8.60086
5	13.05969	16.74781	13.07148	16.81467
6	23.33333	23.13348	10.48025	10.63317
7	18.13558	21.56000	12.56657	7.81385
8	24.02833	32.51292	5.197396	3.31736
9	12.36447	21.47368	12.26675	9.62214
10	18.29765	27.42196	7.569853	5.11276
11	16.53155	21.87534	10.59369	13.14655
12	24.05865	25.10135	8.20926	6.55468

TABLE 8. RESULT OF CALCULATION DIFFERENCE OF IDEAL TEACHER

Cluster	AVG (Equal SMA_ teacher)	AVG (Equal SMA_ Ideal)	Difference of Ideal Teacher	Catagory
1	15.22287	12.19240	3.03048	Enough
2	19.49389	13.15683	6.33706	More
3	11.93816	14.92112	-2.98297	Less
4	21.28698	9.02610	12.26087	More
5	14.90375	14.94308	-0.03932	Less
6	23.23341	10.55671	12.67670	More
7	19.84779	10.19021	9.65758	More
8	28.27063	4.25738	24.01325	More
9	16.91908	10.94444	5.97463	Enough
10	22.85981	6.34131	16.51850	More
11	19.20345	11.87012	7.33333	More
12	24.58000	7.38197	17.19803	More

Based on the calculations in Table 8, the cluster results that have been obtained will be categorized into less categories if the average number of teachers does not exceed the average number of ideal teachers. The results of less category occur in cluster 3 and cluster 5.

The result of the cluster that has been obtained will be categorized into the enough category if the average number of teachers equals the average number of ideal teachers and or has the difference in the average number of teachers with the average number of ideal teachers does not exceed 6.0. The results of enough category occur in clusters 1 and cluster 9.

While the cluster results that have been obtained will be categorized into the more category if the average number of teachers exceeds the average number of teachers ideal and have difference in the average number of teachers and the average number of teachers ideal exceeds 6.0. The results of grouping of more category occur in cluster 2, cluster 4, cluster 6, cluster 7, cluster 8, cluster 10, cluster 11, and cluster 12.

IV. CONCLUSSION

The *k*-means clustering method is well to help group several problems. The distribution of the number of teachers in Indonesia by using the *k*-means clustering method uses a combination of two data, that are the data ratio of school students and data ratio of school teachers in 2015. From the test results, it can be seen that by using cluster with k = 12 has a

higher grouping accuracy as evidenced by SSE results of 87.15%.

Based on the results of the analysis can be obtained that the group is divided into 3 criteria, namely region with less teachers, enough teachers, and more teachers. The region that are less teacher are in cluster 3 and cluster 5 results. Region that are considered to enough teachers are on cluster 1 and cluster 9 results. Whereas the region with more teachers are on cluster 2, cluster 4, cluster 6, cluster 7, cluster 8 , cluster 10, cluster 11, and cluster 12. The results obtained from the difference between the average number of teachers and the average number of ideal teachers.

Based on the calculation of *k*-means clustering, it can be used as a reference to perform the distribution of teachers to region that are less teacher, so as to solve the issue of uneven distribution of teachers.

REFERENCES

[1] Dewey, John, "Democracy and Education", The Free Press, 2009.

[2] Linggasari, Yohannie, "The distribution of teachers is still constrained in local goverment", CNN Indonesia, 2015.

[3] X. Wu and V. Kumar., "The Top Ten Algorithms in Data Mining", Chapman and Hall, 2009.

[4] Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining, (First Edition)", Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc, 2005.

[5] Ong, Johan Oscar, "Implementation of K-Means Clustering Algorithm for Determining Marketing Strategy of President University", Scientific journal of industrial engineering, volume 12, issue 1, 2013.

[6] Narwati, "Student Grouping Using k-means Algorithm", Journal of Informatic Dynamic, volume 2, issue 2, 2010.

[7] Kaur,Navjot et.al, "Efficient k-means clustering algorithm using rangking method in data mining", International Journal of Advanced Research in Computer & Technologi, Volume 1, Issue 3, 2012.