Academic Journal of Engineering and Technology Science
ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

Cloud Computing Refined Data Clustering Algorithm Considering the Generalized Characteristics of Smart Campus

Ruiqing Wang

School of Computer and Information, Anyang Normal University, Anyang 455000, China

Abstract. *with the Rapid Development of Many Technologies, Such as Information Technology, Network Technology, Cloud Computing Technology and Data Technology, the Demand for the Corresponding Analysis and Processing of the Rapidly Growing Data is Becoming High. Data Mining is One of Its Products. in the Process of Data Mining, the Clustering Algorithm is a Very Important Method in the Field of Mining. Therefore, How to Improve the Performance of Clustering Algorithm under Cloud Computing Platform is Very Important.*

Keywords: *Cloud Computing Platform, Big Data, Clustering Mining Algorithm, Parallelization*

1. Introduction

The Purpose of This Paper is to Solve the Problem of the Large Scale Value, It is Faced by Clustering Algorithm by Using the Large-Scale Data Processing Ability of Cloud Computing Platform (Sancho, 2014). This Paper Analyzes the Architecture of Cloud Computing, It Studies the Map Reduce Programming Model and Hdfs Distributed File System, and It Introduces the Related Technologies of Clustering Algorithm. Combining Isodata Algorithm with Map Reduce Programming Model, the Isodata Algorithm Based on Map Reduce is Implemented. We Aim At the Deficiency of Isodata Algorithm, an Improved Algorithm Wisodata is Proposed, and the Wisodata Algorithm Based on Map Reduce is Implemented (Nie,2013). We Select the Well-Known Data Set from the Uci Machine Learning Library, We Analyze and Compare the Clustering Results of the Isodata Algorithm. the Experimental Results Show That the Four Algorithm Clustering Results Have High Accuracy, Wisodata and Wisodata Algorithm Based on Map Reduce Are Both. the Performance of the Isodata and Wisodata Algorithms Based on Map Reduce is Analyzed by Experiments on Different Size Data Sets. the Experimental Results Show That the Isodata and Wisodata Based Algorithms Based on Map Reduce Have Excellent Acceleration Ratio, Data Scalability and Extension Rate (Bosu,2014). They Are Suitable for Running on Cloud Computing Platform, and It Can Be Effectively Applied to Large-Scale Data.

ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

2. Clustering Overview

2.1 The Definition of Clustering

Clustering is an important research area in data mining and statistical analysis, it has attracted much attention in recent years. From the point of machine learning, clustering is an unsupervised machine learning method. That is, first of all, the distribution of data sets (Sarkar, 2015).

Without any understanding, it is a process of composing a collection of physical or abstract objects into multiple classes, which is made up of similar objects, similar to the usual “class of objects”. In many applications, data objects in a cluster can be treated as a whole. Up to now, there has not been a universally accepted definition in clustering. Here is a definition of clustering in 1974: the entities within a class cluster are similar, and the entities of different classes are dissimilar; a class cluster is a convergence of points in the test space, and the distance between any two points of the same class cluster is less than the distance between any two points of a different class of clusters (Pauly,2011).

2.2 The Data Structure Required for Clustering

1) data matrix: It can be seen as a two-dimensional table, each column represents an attribute of the object, each row represents a data object, and this data structure can also be regarded as a matrix of N*M dimension.

$$\begin{vmatrix} P_{11} & P_{12} & \cdots & P_{1m} \\ P_{21} & P_{22} & \cdots & P_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{nm} \end{vmatrix}$$

2) Dissimilarity rectangle: The dissimilarity rectangle stores the dissimilarity between two or more of the n objects. The concrete form is an n*n-dimensional rectangle. Each element d(i,j) is between the object i and the object j. A quantitative representation of dissimilarity is:

$$\begin{vmatrix} 0 & & & \\ d_{21} & 0 & & \\ \vdots & \vdots & \ddots & \\ d_{n1} & d_{n2} & \cdots & 0 \end{vmatrix}$$

ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

2.3 Steps of Clustering

In practical applications, the clustering method used varies with different problems, but a complete clustering process contains the same basic steps, as shown in Figure 1.

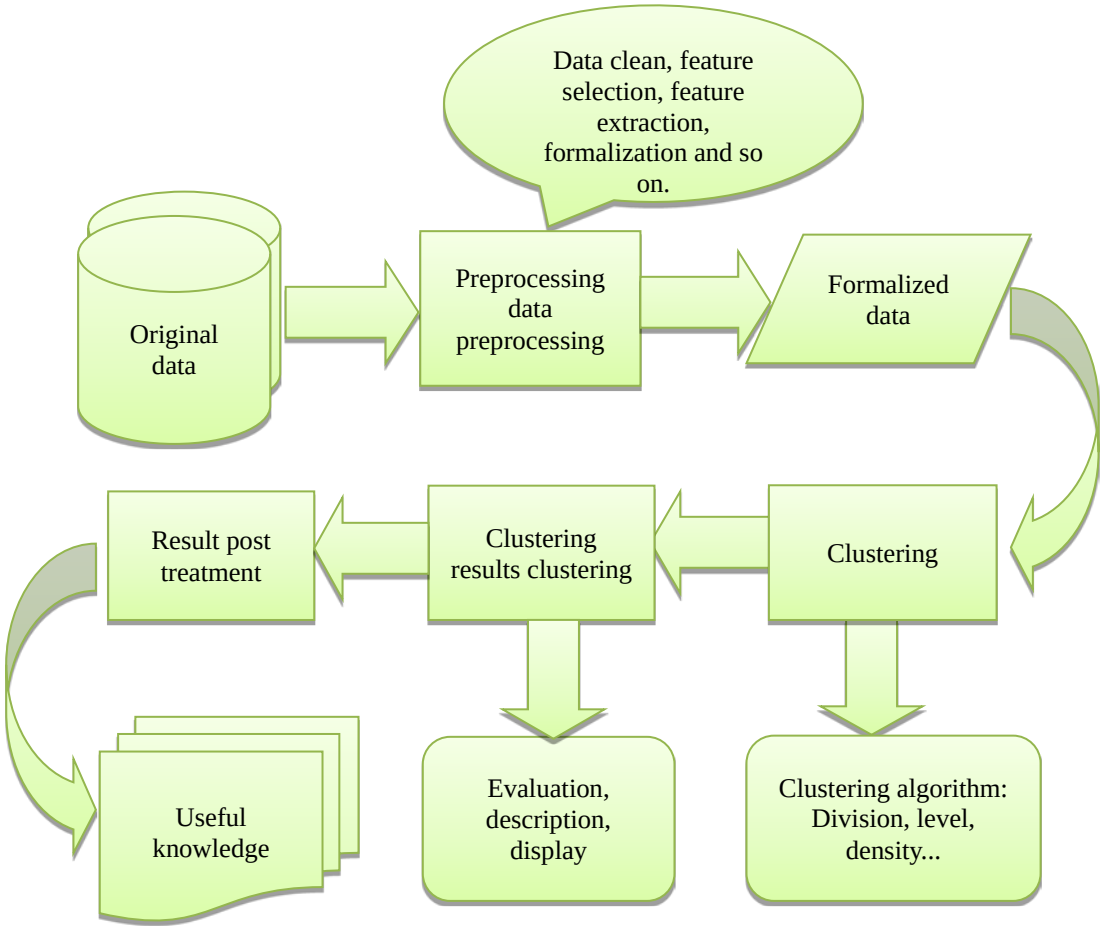

Fig.1 Cluster Process Diagram

3. Cloud Computing and Mapreduce Programming Model.

3.1 Hadoop-Based Cloud Computing Platform

1)Cloud computing technology. Cloud computing is the most reliable and secure data storage center based on the Internet. Users can no longer worry about data loss,

ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

virus intrusion, etc. Cloud computing applications include: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS)(Cantera, 2014).

2)Hadoop. Hadoop is an open source framework for writing and running, it is distributed applications to process large-scale data. Its main feature is that users can develop without understanding the underlying details of distributed systems. Distributed programs make full use of the power of clusters for high-speed computing and storage(Chaczko,2012).

3)Based on the Hadoop cloud computing platform, the HDFS distributed file system is built to store massive text data sets. The distributed index is established through the text word frequency using the Principle of Data Replication, and the distributed database HBase is used to store the index of keywords and provides real-time retrieval. Achieve distributed parallel processing of massive data(Cinellu,2000).

3.2 Big Data Technology

Big Data Technology (Big Data) has become a hot topic of concern, due to the rapid development of the Internet, cloud computing, mobile, and Internet of Things in recent years. Big data is a large-scale acquisition, storage, management, and analysis. The data collection greatly exceeds the capabilities of traditional database software tools. It has a massive amount of data, fast data flow, diverse data types, and low value density. From a technical perspective, the relationship between big data and cloud computing is Indispensable. Big data must use a distributed architecture in its use. Its characteristic lies in the distributed data mining of massive data. But it must rely on cloud computing distributed processing, distributed database and cloud storage, virtualization technology.

3.3 Clustering Mining Technology

Cluster analysis is the core technology used in data mining. Cluster analysis is based on the simple concepts, which are classified according to the characteristics of things. From the recessive level, the results of cluster analysis will be. We call the set of group data objects clusters. The objects in the cluster are similar to each other, but the objects in other clusters are different. In many applications, the data objects in a cluster can be considered as a whole.

Clustering mining technology has been widely studied in fields, such as statistics and artificial intelligence. In recent years, with the development of data mining, Clustering is a very active research topic in the field of data mining, it has unique advantages. In the field of mining, it often faced with a database containing massive amounts of data. Therefore, we must continue to improve the clustering method for large-scale databases to meet the challenges posed by new problems. 2 Improved K-means clustering algorithm research

ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

3.4 Programming Model

Map reduce is a parallel programming model proposed by Google in 2004, which is usually used for programming large data sets, as show in Table 1.

The map reduce framework calculates a set of output <key and value> pairs sets through a set of input keys / values <key and value> pairs. One of its advantages is that it has a high degree of abstraction. In writing map reduce programs, the programmer only needs to pay attention to two functions: the map function and the simplified (reduce) function can be user defined by the map function, which handles an input set of <key, value> pairs, and produces a set of intermediate nodes that are also represented by <key and value>. In the fruit set, the map reduce library aggregates the intermediate value values of the same key, and it sends them to the reduce operation. The reduce function is also user defined, receiving the intermediate key. And its corresponding value set, which sums up these value values and outputs the results.

Table 1 the Functions

function	input	output	Explain
map	(k1,v1)	List(k2,v2)	The data set is further parsed into a batch (key, value) pair, and input the middle result set (K2, V2).
reduce	(k2,list(v2))	(k3,v3)	List (V2) represents the value that belongs to the same K2

Programs written in this way can automatically parallelize on large scale ordinary machines. This allows programmers who have no experience in parallel distributed systems to make use of the resources of a large number of distributed systems.

3.5 Implementation Mechanism

The map reduce framework can have many different implementations. For example, a cluster of hundreds of machines can be built through Hadoop, and a distributed environment and execution effect can be simulated on a single machine. This article will give a detailed description of how to build a Hadoop based cloud environment in the following chapters.

As mentioned above, map reduce is based on the idea of “divide and conquer”, it abstracts computing tasks into two computing processes of map and reduce, which can be simply understood as “decentralized operations - merge results”. A map reduce task first divides the input data into several keys / values for the <key, value> set, which will be handled in parallel by multiple map tasks (a set partition corresponds to a map task). Map Reduce will sort the output of map (a set of

ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

intermediate key value pairs) and combine the values that belong to a key (key, list of values) as the input of the reduce task, and calculate the final result and output by the reduce task. Usually the input and output of the job will be stored in the file system, the entire framework is responsible for the scheduling and monitoring of the task, and the execution of the failed tasks.

4. Isodata Algorithm

4.1 K-Means Clustering Algorithm Basic Ideas and Methods

The K-means algorithm is a typical distance-based clustering algorithm, as show in Figure 2. It uses distance as an indicator of similarity evaluation. That is: If the distance between two objects is closer, the similarity is greater. The algorithm considers clusters to be composed of tightly bound objects. So it is able to get a compact and independent cluster as the final target. The algorithm process is as follows:

1)randomly select K documents from the N documents as the center distance;

2)Measure each remaining document to the center distance, as well as classifying it to the nearest center of mass;

3)Then recalculate the center distance of each class that has been obtained;

4)Iterate the second and third steps until the new center distance and the original center distance are less than or equal to the specified threshold;

5)The algorithm is all over. The specific algorithm is described as follows: Input: k, data[n];

6)Select k initial center points, for example c[0]=data[0], c[k -1]= data[k -1];

7)For data[0]... .data [n], respectively, compared with c[0]...c[k -1], first assuming that it has the least difference with c [i], the label is i;

8)Pairs of all marked i points, and then recalculated c[i] ={sum of all data [j] marked i).

ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

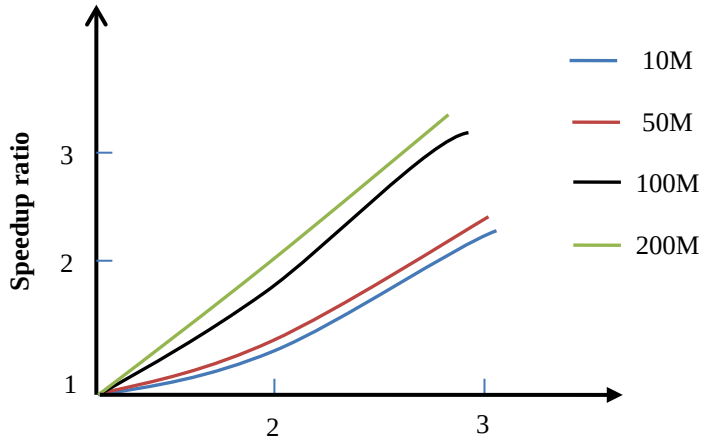

Fig.2 K-Means Algorithm Operation Acceleration Ratio

4.2 K-Means Algorithm

The K- means algorithm is usually suitable for classifying the known number of clusters, while the ISODATA algorithm is more flexible, as show in Figure 3.

From the algorithm point of view, the ISODATA algorithm is similar to the K-mean algorithm, and the cluster centers are determined by the iterative operation of the sample mean. The ISODATA algorithm adds some exploratory steps, it can combine the structure of adult computer interaction so that it can be better classified by the experience and it is obtained by the intermediate results.

ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

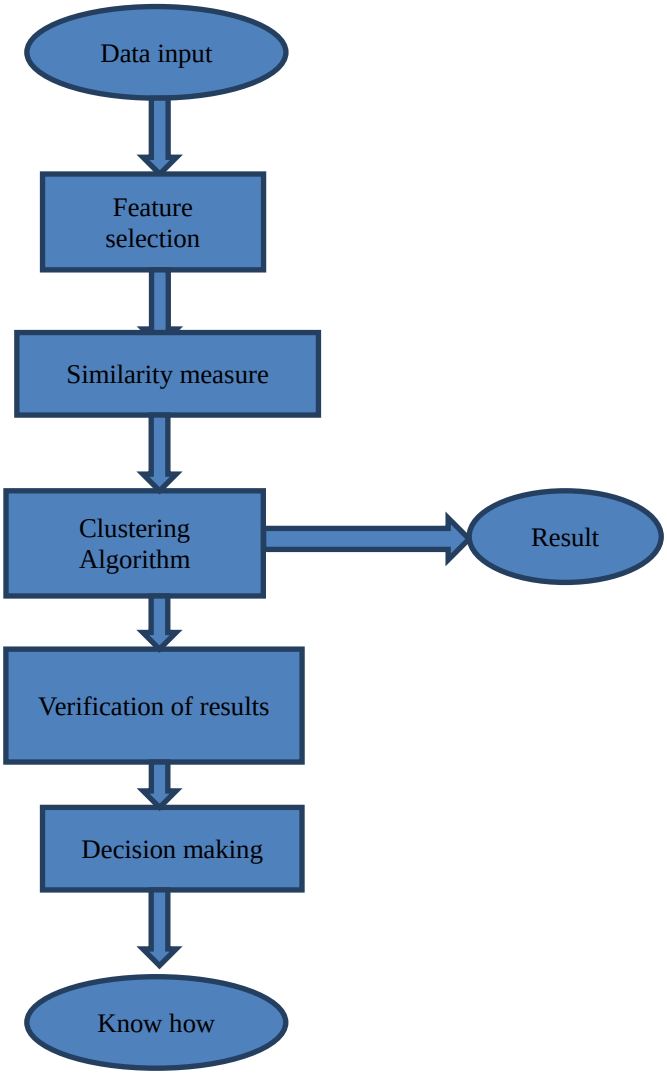

Fig.3 General Flow of Clustering

4.3 Example:

ISODATA algorithm is used to cluster analysis of the following pattern distribution:

$$\{X_1(0,0), X_2(3,8), X_3(2,2), X_4(1,1), X_5(5,3), X_6(4,8)\}$$

ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

The implementation of clustering algorithm on Hadoop platform is complex. It needs to implement the map and reduce functions. The implementation of the DBIK-means clustering algorithm on Hadoop is divided into two phases. The main content of the design is to design and implement the map. And R educe functions, including the types of input and output <key, value> key-value pairs, and the specific logic of map and R educe functions, as show in Table 2.

Table 2 Comparison of Algorithm Performance

Sample point	characteristic value		The distance to Z1	The distance to Z2	clustering results
X1	0	0	6.3893	4.6537	S2
X2	3	8	3.0737	5.5347	S1
X3	2	2	3.6027	1.975	S2
X4	1	1	4.9902	3.2831	S2
X5	5	3	2.3362	1.1927	S2
X6	4	8	2.9407	5.4619	S1
X7	6	3	2.9424	2.15	S2
X8	5	4	1.5283	1.8288	S1
X9	6	4	2.3528	2.5582	S1
X10	7	5	3.1006	3.9581	S1

Map function design input map function comprises a set of <key, value> key set and the density threshold min P ts pairs, where <key, value> is map reduce frame default format, key represents The current dataset offset relative to the starting point of the input data file, value is the corresponding data; the output of the function is a set of <key 1, value 1> key value pairs, where key1 represents the cluster number, and value 1 represents The related attributes belonging to the key1 cluster, including the center point of the cluster, the number of cluster midpoints N um, the sum S um of all points in the cluster, and the degree of dissimilarity between the point farthest from the center point in the cluster and the center point , i is in the range of 1 to k, and k is the number of clusters, as show in Table 3.

Table 3 Comparison of Algorithm Clustering Results

Sample point	Characteristic value		The distance to Z1	The distance to Z2	Clustering results
X1	0	0	7.6577	3.3287	S2
X2	3	8	2.9732	6.2032	S1
X3	2	2	4.8415	0.82462	S2
X4	1	1	6.2482	1.9698	S2
X5	5	3	2.8	2.506	S2
X6	4	8	2.4166	6.3151	S1
X7	6	3	2.9732	3.4176	S1
X8	5	4	1.8	3.1113	S1

ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

X9	6	4	2.0591	3.8833	S1
X10	7	5	2.1541	5.2802	S1

Big data technology, Internet of things, and cloud computing are disruptive technological revolutions in today's IT industry. The arrival of the era of big data has had an important impact on people's lifestyles and business models. Big data has been proposed to the IT industry belt. New development opportunities have emerged, especially for data mining technologies. Data mining technology has entered a new stage of development. To improve the accuracy of big data, big data mining algorithms have large error rates for large data processing. It must be in an acceptable range, which requires the improvement of traditional data mining algorithms, as show in Table 4. The article aims to improve the efficiency and accuracy of big data mining, and focuses on the accuracy and efficiency of clustering algorithms for big data. The traditional clustering algorithm makes necessary improvements, and we use the cloud computing platform to parallelize the improved clustering algorithm. The improved clustering algorithm has good theoretical and practical value, and it can be used in large data sets. The platform is widely promoted and applied.

Table 4 Comparison of Algorithm Clustering Results

Sample point	Characteristic value		The distance to Z1	The distance to Z2	The distance to Z3	Clustering results
X1	0	0	8.1626	6.7523	2.5	S2
X2	3	8	2.7421	4.247	6.5765	S11
X3	2	2	5.3558	3.9418	0.5	S2
X4	1	1	6.7569	5.3447	1.118	S2
X5	5	3	3.3235	1.3576	3.3541	S12
X6	4	8	2.046	3.8345	6.8007	S11
X7	6	3	3.4223	1.5842	4.272	S12
X8	5	4	2.3253	0.38524	3.9051	S12
X9	6	4	2.4645	0.90278	4.717	S12
X10	7	5	2.2587	1.946	6.1033	S12

4.4 Density-Based Incremental Dbik-Means Improvement Methods K-Means Algorithm

Disadvantages: In the K-means algorithm, K is often given, so the selection of K values is difficult to estimate. Before the operation, it is unclear how many categories a given data set should be divided into. It is most appropriate to use the K-means algorithm to determine the initial partition based on the initial cluster center and then to optimize it, as show in Table 5. This choice will produce results

ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

for the cluster analysis. For larger impacts, if the initial values are not well selected, it may not be possible to obtain effective clustering results. In addition, it can be found in the framework of the k-means algorithm, which takes a long time when the amount of data to be manipulated is large. Therefore, it is necessary to improve the algorithm time complexity to a certain degree.

Combining the disadvantages of k-means clustering algorithm and the basic idea of density increase, we propose a density based incremental k-means clustering algorithm.

Table 5 Comparison of Algorithm Performance

Data sets and files	K-means (correct rate %)	LAFSA-KM (correct rate %)	Distributed LAFSA-KM algorithm (correct rate %)
date1	90.1	94.3	95.7
Date2	71.1	86.2	89.4
Date3	83.2	89.3	90.3

The article uses open source distributed software Hadoop, it is build to experimental cloud computing platform for testing the performance of the DBIK-means clustering algorithm. A total of four computers were used for testing and analysis, the operating system was installed fedora 9, a machine is master. The remaining three machine slaves, when testing the hdfs distributed storage, data backup into three copies, when testing closed one of the slave machines to observe whether the overall file system operation is affected, as show in Table 6 and Table 7.

Table 6 Validity and Timeliness of the Algorithm for Cure under the Stand

algorithm		CURE algorithm		
data set	Iris	Wine	Libras	Diabetes
Correct rate (%)	80.1	94.4	44.2	74.3
Running time (s)	18.2	11.58	18.33	23.12

Table 7 Effectiveness and Timeliness of Ms-Cure's Algorithm

algorithm		MS-CURE algorithm		
data set	Iris	Wine	Libras	Diabetes
Correct rate (%)	82.3	95.8	46.3	74.3
Running time (s)	2.68	2.78	2.7	3.8

5. Conclusion

ISSN 2616-5667 Vol.2, Issue 5: 45-56, DOI: 10.25236/AJETS.2019.020506

Cloud computing is the product of the development of computer science and technology to a certain extent. In the current network technology, cloud computing is a new computing model, and it is also the key technology in the next generation network computing platform. The limitations of the traditional campus network environment mainly contain the large energy consumption, poor performance and resource benefits. With the lack of other aspects, cloud computing can effectively integrate the resources, and greatly reduce energy consumption and ensure the security of information. So the construction of intelligent campus should be the core of the cloud computing data center.

References

[1] Sancho-Asensio, A., Navarro, J., Arrieta-Salinas, I., et al.(2014). Improving data partition schemes in smart grids via clustering data streams. Expert Systems with Applications, vol. 13, no. 41, pp. 5832-5842.

[2] Nie, X(2013). Research on smart campus based on cloud computing and internet of things. Applied Mechanics & Materials, vol. 1, no. 380-384, pp. 1951-1954.

[3] Bosu, A., Carver, J. C., Hafiz, M., et al.(2014) Identifying the characteristics of vulnerable code changes: an empirical study. ACM Sigsoft International Symposium on Foundations of Software Engineering, vol. 1, no. 1, pp.257-268.

[4] Sarkar, S., Misra, S., Bandyopadhyay, B., et al.(2015) Performance analysis of ieee 802.15.6 mac protocol under non-ideal channel conditions and saturated traffic regime. IEEE Transactions on Computers, vol. 10, no. 64, pp. 2912-2925.

[5] Pauly, D., Rossi, T.(2011). Theoretical considerations on the computation of generalized time-periodic waves. Mathematics, vol. 6, no. 5, pp. 302 – 311.

[6] Cantera, M. A., Romera, J. M., Adarraga, I., Mujika, F. (2014). Modelling of laminates subject to thermal effects considering mechanical curvature and through-the-thickness strain." Composite Structures, vol. 110, no. 110, pp. 77-87.

[7] Chaczko, Z., Aslanzadeh, S., Kuleff, J. (2012). The artificial immune system approach for smart air-conditioning control. International Journal of Electronics & Telecommunications, vol. 2, no. 58, pp. 193-199.

[8] Cinellu, M. A., Minghetti, G., Pinna, M. V., et al. (2010).Organometallic Gold(Ⅲ)Derivatives with Anionic Oxygen Ligands-mononuclear Hydroxo,Alkoxo,and Acetato Complexes：Synthesis and Spectral Study. Journal of the Chemical Society Dalton Transactions, vol. 8, no. 41, pp. 1261-1265.