IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017), PP 114-121
www.iosrjournals.org

Hadoop Based Big Data Clustering using Genetic & K-Means Algorithm

Palak Sachar[1], Vikas Khullar[2]
[1]*(Student of Masters of Technology in Computer Science and Engineering)*
[2]*(Assistant Professor in Computer Science and Engineering)*
CT Group of Institute, Jalandhar, India.

Abstract : *This is the era of huge and large sets of data or can say Big Data. Clustering of Big data plays several important roles for Big Data analytics. In this paper, we are introducing Big Data clustering algorithm by combining Genetic and K-Means algorithm using Hadoop framework. The major aim of this hybrid algorithm is to make clustering process faster and also raise the accuracy of resultant clusters.*

Keywords Big Data Analytics, Genetic Algorithm, Hadoop, K-Means, MapReduce,

.

I. Introduction

Genetic algorithm is famous for optimization and K-Means algorithmis one of the best for data clustering. These techniques are able to obtain optimal results in global search space and produces outcomes in less time respectively [1,2]. Paper [3] is panoply of working of Genetic Algorithm for clustering data. Paper exhibits a comparison of parallel algorithm and sequential process in which Genetic Algorithm on Hadoop platform overpower the other in respect of Time and accuracy. These algorithms also had lacunas such as the requirement of prior detail knowledge of the various input parameters and K-means clustering algorithm get early convergence perspicacity. This prior knowledge requirement is essential to get desired outcomes and early convergence perspicacityis not correct according to algorithm requirement lead to poor results. Genetic and K-Means algorithm best fitted together for avoiding these problems during clustering to get optimal solution in lesser time.In present scenario Big Data made directed different techniques towards itself for handling lager Volume, higher Velocity and Variety of generated data. In this paper, we proposed, implemented and analyzed a hybrid approach of optimized clustering using Genetic and K-Means algorithm with the support of Hadoop MapReduce along with Mahout clustering libraries [3].

There are many more benefits as well which validate to endeavor.Social media is the only source to get large genuine data where we can verify our approach. So, we choose to work on Twitter based data as it is well suited to verify any approach. It is difficult to collect lest we have a great internet connection or superb patience. [4]

II. Background

a. Clustering Techniques

In Data Mining, there are numerous Clustering Techniques out of which K-mean clustering is an important technique because of the computations being able to revolve around the user defined centroids and complete its task in no time [1, 14]. Another important clustering Technique is called hierarchal clusters which make the tree like structures. It divides the huge cluster into smaller ones based on similarities until it gets to the k-clusters. Other clustering techniques are DB scan and optical clustering techniques which are based on density theory of data. Optical Algorithm is superior to the former one because it overcomes the limitations of finding more relevant data into the clusters from the data having least dense characteristics [1, 10]. Apart from these, there is graph theory clustering techniques which helps in computing the clustering of data through graphs. Model-based clustering techniques involve Decision trees and networks. In 2001, Grid clustering techniques like fuzzy logics and an evolutionary algorithm were introduced. In a study by Akilesh and his team (2015), it was shown that a particle swarm optimization algorithm is better than K-means algorithm onto the Map-Reduce framework [6]. Similarly, the Genetic algorithm is a unique and growing technology in the research field [2]. Cellular genetic algorithm (7) was also used to do clustering of tweets and behaviour was seen on JAVA platform [18].In paper[3], one of the Evolutionary Algorithmi.e. a classic Genetic Algorithm has been pre-owned to do Clustering of Big Data on Hadoop platform which leads to give a better execution Time and accuracy as compare on JAVA platform.

\

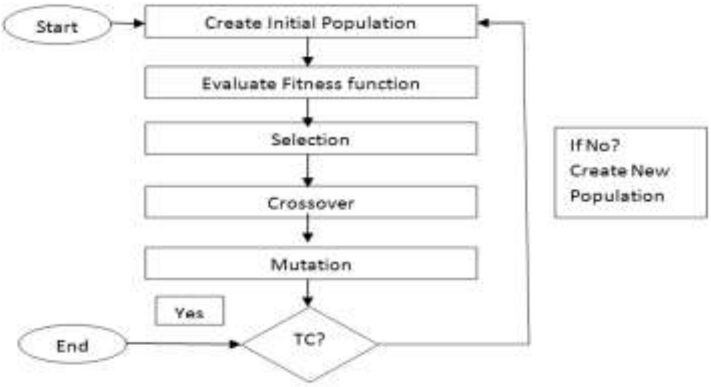

Figure 1 Procedure of Genetic Algorithm [2]

b. Genetic Algorithm

Genetic Algorithm is a metaheuristic approach capable of producing an optimized solution. Darwin's theory of evolution based genetic algorithm has proven to be a stepping stone in the field of data mining, medicine and data science. Genetic algorithm is an adaptive procedure. Figure 1 represents the all five series of steps in GAs. Genetic algorithm is a series of five steps which includes initialization; evaluate fitness function and selection; crossover; mutation and termination criteria as shown in Figure 1. Genetic Algorithm varies with the definition of problem. The main aspect of genetic algorithm is based on three questions - (1) How to initialize the population? (2)How to evaluate fitness function? (3) What would be the termination point? It defines the layout of GA's efficiency [3, 13]. Genetic Algorithm which is used for Clustering is shown in Figures 3 and 4. These figures depict the working of Algorithm on Hadoop MapReduce.In figure 3, it shows the main program and figure 4 is the MapReduce part of the program which exhibits the basic Genetic Algorithm steps.The next challenge in the research is to make accuracy better without affecting the Cluster Data and even to make algorithm robust in such a way that it can handle more data like millions of Data. So we decided to hybrid it with another technology.

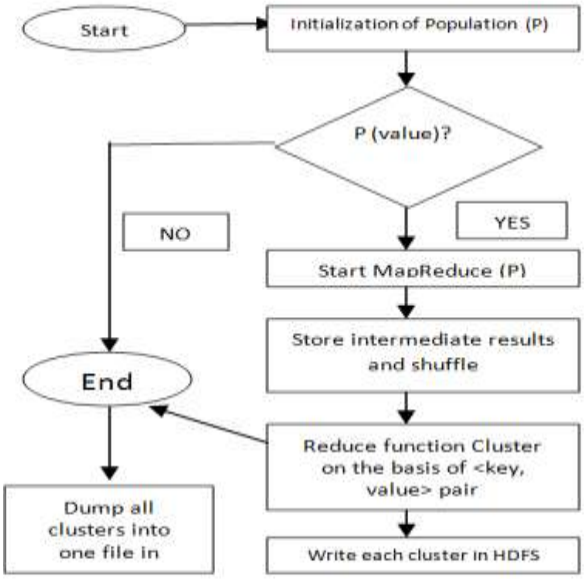
Start
Initialization of Population (P)
P (value)?
NO
YES
Start MapReduce (P)
Store intermediate results and shuffle
End
Reduce function Cluster on the basis of <key, value> pair
Dump all clusters into one file in
Write each cluster in HDFS

Figure 2 Genetic Algorithm for Clustering (A)

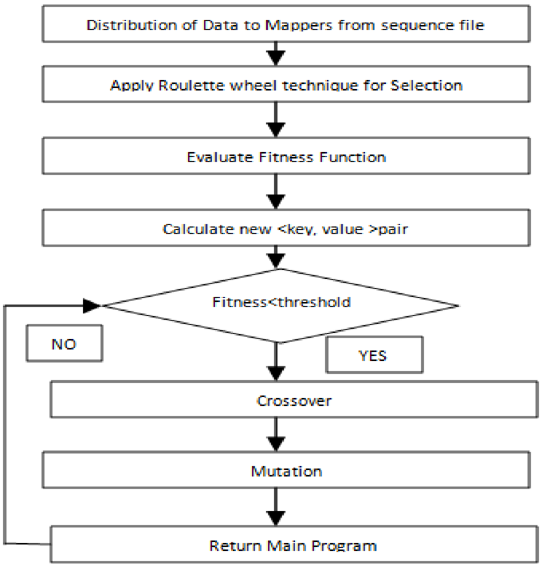
Distribution of Data to Mappers from sequence file
Apply Roulette wheel technique for Selection
Evaluate Fitness Function
Calculate new <key, value >pair
Fitness<threshold
NO
YES
Crossover
Mutation
Return Main Program

Figure 3 Genetic Algorithm for Clustering (B)

c. Hadoop MapReduce

MapReduce evaluates data sets in two phases -mapped phase and reducer phase also shown in Figure 2. Both phases determine their respective functions - map function and reduce function. It follows the strategy of divide and conquer [20].Mapped phase splits the large data sets associated with map function and store results in an intermediate stage with the key, value pair. A key is randomly generated unique number while value has a computed value defined by the programmer. From the intermediate stage, data gets shuffled and reducer gets data according to their corresponding key and value. Reduce function calculates the correlated function and produces output for the job by writing it into HDFS. [1, 11, 12]

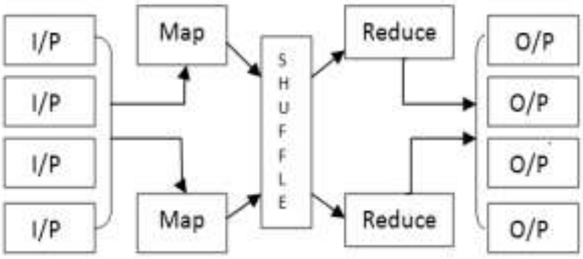

Figure 4 Working of MapReduce [3]

d. Mahout

Mahout is an arrangement of machine learning. JAVA libraries are meant for various tasks such as classification, evaluation clustering, and pattern mining and so on. Mahout framework is capable of handling large datasets. The power of mahout lies in the fact that the algorithms are used in a Hadoop environment. Collaboration with Hadoop permits for an algorithm to run in parallel using distributed computing paradigm. [20]

e. Twitter API

Twitter is a microblogging site that produces s large amount of data in form of tweets. Twitter is a powerful social networking site which provides application interface for developers [8]. There are two API's - rest API and streaming API. Rest API shows all the tweets on a twitter server. Streaming API is more robust, generic and is user oriented. For streaming API, the user has to create an application on Twitter Development site. Twitter provides four unique and confidential tokens for each user which establishes the connection between user and twitter servers. It’s also useful in verification of the valid user. One has to create it with the platform either using java or python. To collect a large amount of data, a strong internet connection is needed. The tweets can be collected or stored in any storage format. In the present study, CSV files are used to store tweets.Collection of tweets was done using Twitter API. Next step is removing stop words as it has zero relevance in clusters. Stemming is a procedure of removing forms of verbs and adverbs and categorizing them. For example: sleeping, slept will be considered as sleep category.It is followed by tokenization in which text is divided into tokens (A word is considered as a token). Pre-processing of tweets is shown in Figure 2.

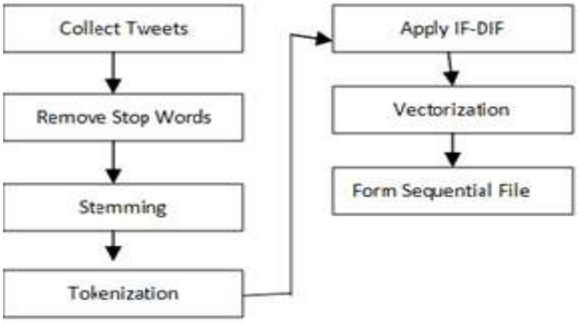

Figure 5 Pre-Processing of Tweets [6]

TF-IDF is applied to evaluate the value of each token followed by vectorization to form a sequence file. Vectorization is a process of converting tokens into 0s and 1s by evaluating its key value pair on the distributed paradigm. [15, 19]

III. Methodology

Whole work is divided into three phases are as follows:
1. Collection of Tweets via Twitter API and its Preprocessing.
2. Generate a Vector File using Genetic Algorithm.
3. Run K-Means Clustering Algorithm.
Planned exposition of above phases is as follow:

A) Collection of Tweets

Collection of Tweets is done using Twitter API with third party python based application. Twitter provides two types of API which can gather Data in their own way. One is to collect static Data and another one is live Data to congregate Tweets using Static API and Stream API respectively. The two requires an essential parameter represent by letter ‘q’ around which whole search revolves in the tweets servers. Data amass is dependent on internet speed. With the great bandwidth, user is able to collect data with the range of 50k-80k tweets.

The collection is full of noise and lot of information it consist. It is very important to eliminate unwanted noise and to be deducted from the dataset. For this purpose, one needs to apply porter and stemmer algorithm which remove whole noise from the dataset.

B) Generation of Vectors using Genetic Algorithm

The milestone of entire work process is to reduce time as it is the most important factor in today's research. So to cut time in real manner, we do vectorization of Tweets using Genetic Algorithm. The following algorithm is capable of producing a pair of <key, value> which is integral part and responsible of entire clustering computations. The Preprocessing of tweets leads us to a vector file may called a sequential file consist of 0s and 1s.The input of this sequential file is a dictionary file which work as key responsible for making perfect Sequential file and later to convert output file into English like readable files.

Algorithm:-

```
1. Create initial random set
2. Loop
   a. evaluate the whole set of rules against the    main dataset
   b. calculate the fitness function
   c. generate new set of rules
   d. until fitness < threshold
```

C) K-Means Clustering Algorithm

K-Means Clustering Algorithm is best for various reasons. But after the computations of <key, value> pair in the previous stages and centric computations, it makes the working of whole program even better.

```
1. Initialization of Population
2. while (P has next value)
3. Run Gave and save results in GA_SF
4. Map_reduce=start(Mapreduce_job(GA_SF)
5. P=map_reduce(p)
6. Save results in intermediate stage.
7. Reducer will start clustering according to <key, value> pair save into HDFS.
8. Step 3 repeat till all intermediate stage has values.
9. HDFS write all clusters into one file.
```

MapReduce (GA_SF)

```
1. Splitting the sequence file into Mappers.
2. Loop ( _threshold)
3. Mappers evaluate <key, value> pair and respective centroid.
4. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the
   cluster centers.
5. List(evol_op) <rule> op = new arrayList(evol_op) <new rules>;
6. Recalculate the distance between each data point and new obtained cluster centers.
7. Evol_Pipeline pipeline = new EvolutionPipelin(op);
8. If no data point was reassigned then stop, otherwise repeat from step 3
```

Results were calculated on proposed approach, evaluated on above mentioned parameters and compared with Genetic Algorithm standalone approach on Hadoop MapReduce. Table 2 and Table 3 perpend the numerical values observed during the experiment.

D) Parameters

Following parameter has been used in the research to evaluate the performance.

1. Execution Time
2. Accuracy
3. Number Of Iterations
4. CPU Time Spent
5. Total Heap Committed

The detail discussions of the parameter are as following:

Execution Time (T) is the total time taken by process. It is difference between the times at which final cluster (C_f(t)) dump to the Start Time(s(t)).

$$T=C_f\ (t)-S\ (t)\ldots\ldots\ldots\ldots\ldots\ (\text{eq. } 1)$$

The execution Time is calculated in minutes.

Accuracy (A) is the percentage of label occurred most frequent in the output file to the total number of labels.

Accuracy (%) =(Label which occurred most frequent/Total Number of Labels)*100 ...(eq. 2)

The output file consists of clusters plus adjacent labels which tells the purity of particular tweet in the entire cluster. So Accuracy may know as purity of the Tweets.

Number of iterations (n) tells the iterative structure in which the n is quantum number represents how many iterations a program may take to complete the task. More the number iterations, more will be time consumed, more will be the cluster cost. So it is important to make program more reliable at the user end, iteration must be cut.

Number of iterations (n) = level at which final cluster is received…… (eq. 3)

The program shows the number of loops process has been taking to complete the task. The final number (n) is recorded on the Hadoop console.

Moreover, number of iterations also affects the time complexity of the program. Since K-Means Clustering Algorithm have O (k*n*c), where k is the centroid, n is the number of iterations and c is the number of clusters have been written.

CPU Time spent (Time$_{\mathbf{CPU}}$) is the total time utilize by the CPU for the allocation of resources like Input/output devices and much more. It is calculated in Millisecond.

Memory Utilization (M) is the numbers of bytes have been written during clustering the Tweets. It is calculated in bytes and percentage is found out with Maximum Heap Size for the ease.

Memory (%) = (Total Heap Committed/ Max. Heap Size)*100 . … (eq. 4)

E) Tweets Collection

Total Tweets were collected 1 million and results have been taken by dividing the dataset of 1 million evenly into 5 cases.

F) Machine Configuration

Machine configuration is shown in following Table.

Table 1 Machine configuration

Operating System	Ubuntu 12.04 LTS
Java Version	Oracle-6-Java
Hadoop Version	Hadoop 1.2.
File System	Hadoop Distributed File System
Mahout	0.6
Maven	3.0.4
System	Node 1
Processor	Intel Pentium Processor
Hard Disk	250 GB
RAM	3 GB

User may use upgraded system and even on the cluster of machines.

IV. Results

Results for proposed approach were calculated and evaluated on above mentioned parameters. The calculated results compared with Genetic algorithm based clustering technique on Hadoop framework.

Execution Time is evaluated in minutes as a unit of measurement. Figure 3and table 2 reflected the results for ten lakh Tweets. Here recorded execution time of Genetic Algorithm lies between 30 to 244 minutes where as our proposed algorithm resulted between 13 to 175 minutes approximately. Approximately 34 % of changes had noted for 10 lakh of tweets.

Table 2 Results of Proposed Algorithm with comparison of Genetic Algorithm on parameter-Execution Time.

Tweets (lakhs)	Execution Time(minutes)	
	Proposed Algorithm	Genetic Algorithm
2	13.58	30
4	40.21	47
6	52.52	60.01
8	81.21	90.73
10	174.29	244

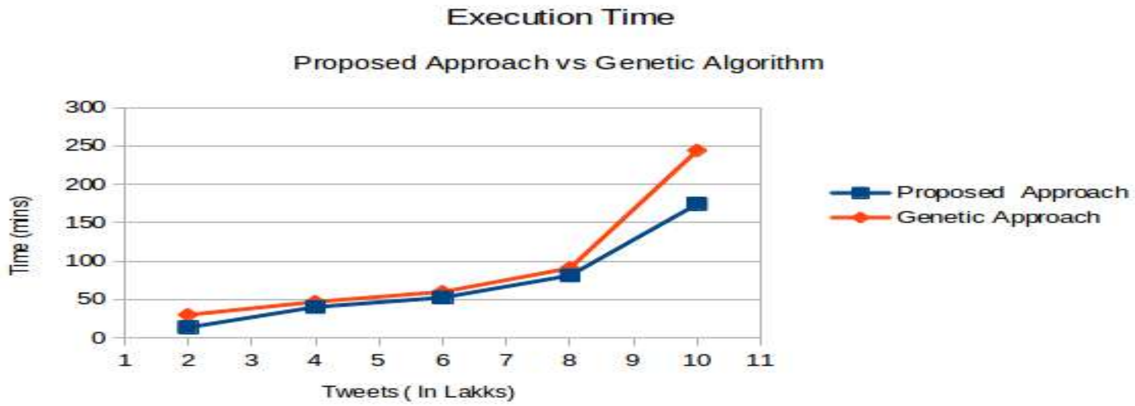

Figure 3 Graphical layout of Results on the basis of Execution Time

Number of Iterations (shown in figure 4 and table 3) is increasing with the increment in volume of data sets for both algorithms. The difference of number of iterations in Proposed and Genetic algorithm approach is minor but the difference shows the positive results with the constant increase in size of dataset.

Table 3 Results of Proposed Algorithm with comparison of Genetic Algorithm on parameter Number of Iterations.

Tweets (lakhs)	Number of Iterations	
	Proposed Algorithm	Genetic Algorithm
2	11	12
4	16	18
6	17	20
8	20	22
10	25	27

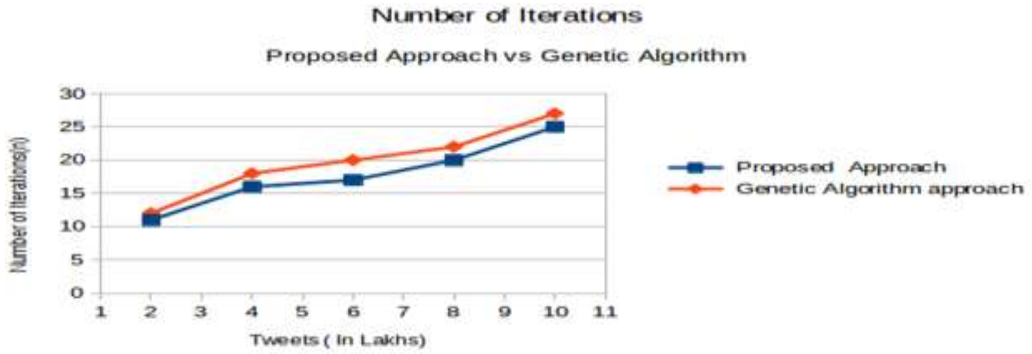

Figure 4 Result: Number of Iterations

Accuracy (in figure 5 and table 4) depicted the stable values in both the cases. We also had noted better accuracy in case of proposed technique as compare to already available technique.

Table 4 Results of Proposed Algorithm with comparison of Genetic Algorithm on parameter –Accuracy

Tweets (lakhs)	Accuracy (%)	
	Proposed Algorithm	Genetic Algorithm
2	93	89
4	92	90
6	93	91
8	94	92
10	94	93

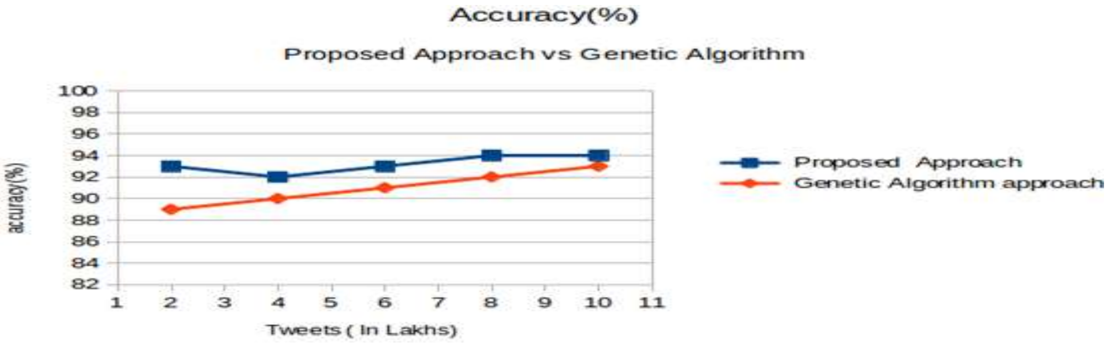

Figure 5 Result: Accuracy

Memory Utilization (in figure 6 and table 5) in the proposed approach is less than the Genetic algorithm approach which comprehends in the manner that proposed technique is consuming less memory or it have less memory requirements than the already available technique.

Table 5 Results of Proposed Algorithm with comparison of Genetic Algorithm on parameter Memory Utilization

Tweets (lakhs)	Memory Utilization (%)	
	Proposed Approach	Genetic Algorithm
2	6.608	37.09
4	5.45	38
6	10.66	38
8	44.39	59
10	58.03	80

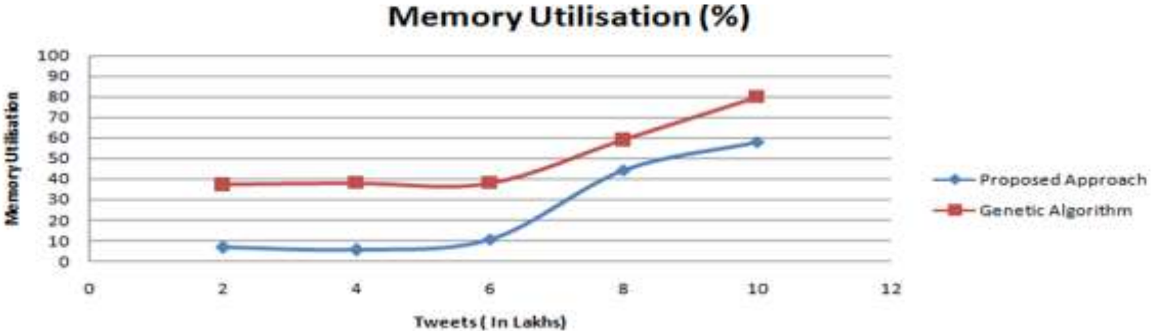

Figure 6 Comparison of Memory Utilization

CPU Time (in figure 6 and table 5) in the proposed approach is much lesser than the Genetic algorithm approach. It reveals the requirement of CPU processing time or complexity is quite low in proposed algorithm.

Table 6 Results of Proposed Algorithm with comparison of Genetic Algorithm on parameter CPU Time

Data(in lakhs)	CPU Time Spent(ms)	
	Proposed Algorithm	Genetic Algorithm Approach
2	44150	229890
4	93840	240920
6	146450	407632
8	209020	809720
10	961450	1392640

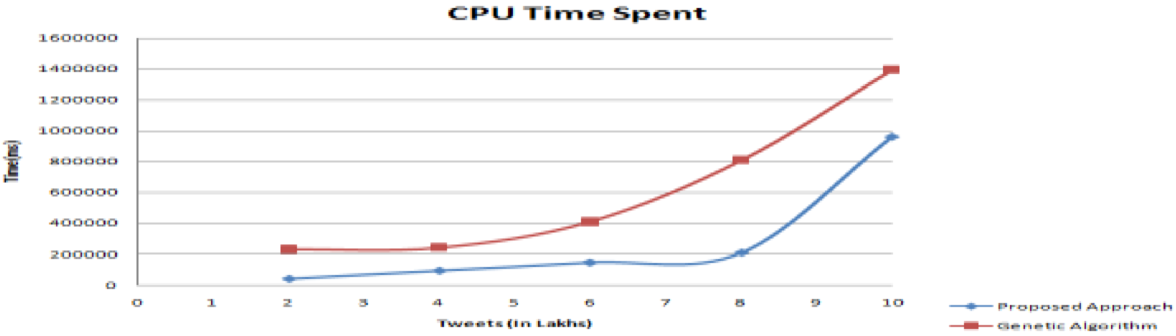

Figure 7 Result: CPU Time Spent

V. Conclusion and Future Scope

It is concluded that the proposed approach is beneficial in the terms of overall execution and CPU time. It had also provided better accuracy and required lesser iterations, which paramount the qualities of clustering algorithm. Memory utilization by our proposed algorithm led us to conclude that memory cost is much lesser than Genetic algorithm based clustering. By using better configuration machines we can study more volume data sets on other clustering algorithms.

References

[1] Amr Adel, Esaam ElFakharany and Amr Badr, “Clustering Tweets Using Cellular Genetic Algorithm”, Journal of Computer Science, Volume 10, Issue 7, pp 1269-1280, 2014.

[2] Filomena Ferrucci, M-Tahar Kechadi, ”A Parallel Genetic Algorithm Framework Based on MapReduce”,30th annual ACM symposium on Applied Computing, pp 1785-1793, 2015.

[3] Palak Sachar and Vikas Khullar,”Social Media Generated Big Data Clustering Using Genetic Algorithm”, International Conference of computer Communication and informatics (IEEE), 2017.

[4] Aditya B. Patel, Manashvi Birla, Ushma Nair, “Addressing Big Data Problem using Hadoop and MapReduce”, Nirma University International Conference on Engineering (IEEE) , 2012

[5] Isabel Anger and Christian Kittl, “Measuring Influence on Twitter”, 11th international Conference on Knowledge Management and Knowledge Technologies (ACM), 2011.

[6] Akilesh P. chunne , Uddagiri Chandrasekhar , Chetan Malhotra , “ Real Time Clustering of Tweets using Adaptive PSO Technique and MapReduce”, Global conference of Communication Technologies (IEEE) , pp 452 – 457 , 2015

[7] Chengyu Hu,Jing Zhao,Xuesong Yan ,Deze zeng and song guo,”A MapReduce based Parallel Niche Genetic Algorithm for containment Source identification in Water Distributed Network “.Journal of Ad Hoc Networks, volume 4 , pp 1-11,2015.

[8] Amin Mohebi , S.R Aghabozorgi, Teh Ying Wah ,T Herawan and R Yahyapour , “Iterative Big Data Clustering Algorithms-A review”, Journal of Software and Practical Experience, volume 46 , pp 107–129 , 2016.

[9] Atanas Radenski, “Distributed Simulating Annealing with Map Reduce”, European Conference on Applications of Evolutionary Computation (Springer), pp 466-476, 2012.

[10] Java A, X. Song, T. Finin and B. Tseng,”Why we Twitter: Understanding Microblogging Usage and Communitie”, Proceedings of the 9th Workshop on Web Mining and Social Network Analysis (ACM), pp 56-65, 2007.

[11] Periasamy Vivekanandan, and Raju Nedunchezhian, (2011). “Mining Data Stream with the Concept Drifts using Genetic Algorithm”, journal of Springer Science, Volume-36, pp 163-171, 2011.

[12] Abhishek Verma , X Llorà , D.E Goldberg , Roy H.Campbell,” Scaling Genetic Algorithms Using MapReduce”, on 9th International Conference on Intelligent Systems Design and Applications (IEEE) ,pp 13-18 , 2009.

[13] Fei Teng, and Doga tuncay, "Genetic Algorithm using MapReduce runtimes”, proposal document on salsahpc.indiana.edu.

[14] Abdelhak Bousbaci and Nadjet Kamel,”A parallel sampling-PSO-multi-core-Kmeans algorithm using MapReduce”, published in 14 international conference of Hybrid intelligent system (IEEE), 2016.

[15] Ian Davidson and Ashwin Satyanarayn ,”Speeding up K-means Clustering by bootstrap averaging” published in 3rd international conference on Data Mining ,Workshop on Clustering Large Data Sets, pp 16-25,2003.

[16] Robson L.F.Cordeiro , Caetano Traina Junior , Agma J.M. Traina ,”Clustering very large multi-dimensional dataset with MapReduce ”Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining , pp 690 -698 , 2011.

[17] Khaled M.Hammouda and Mahamed S. Kamel,”Hierarchically Distributed Peer-to-Peer Document Clustering and Cluster Summarization”, published in Transactions on Knowledge and Data Engineering (IEEE), 2009.

[18] Santhana Chaimontree , Katie Atkinson and Frans Cornen ,”A Framework for Multi-Agent Based Clustering”, published in journal of Autonomous Agents and Multi-Agents Systems, Volume 25, Issue 3 ,pp 424-446,2012.

[19] Johann M kraus and Hans A Kestler, ”A Highly Efficient Multi-Core Algorithm For Clustering Extremely Large Datasets”, published in journal of BMC bioinformatics, Volume 11,pp 169 – 177,2010.

[20] Surasith Taokok,Prach Pongpanich, Nittaya Kerdprasop and Kittisak Kerdprasop ,”A Multi-Threading In Prolog To Implement K-Mean Clustering”, published in Latest Advances in systems Science and Computational intelligence, pp 120-126,2012.

[21] Nandjet Kamel, Imane Ouchen and Karim Baali, “A Sampling-PSO-K-means Algorithm for Document Clustering”, published in journal of Genetic and Evolutionary Computing (Springer), pp 238-246, 2014.

[22] Xiaohi Cui and Thomas E.Potok , ”Document Clustering using Particle Swarm Optimization” published in Swarm intelligent Symposium(IEEE),pp 220- 225,2005.

[23] Palak Sachar and Vikas Khullar , “ Genetic Algorithm using MapReduce-A critical Review”, i-manager’s Journal on Cloud Computing, Volume 2, No 4, 2016.