QUALITY AND ACCURACY OF CLUSTERING ALGORITHMS ON BIG DATA SETS USING HADOOP

Chandrahas Reddy.M\textsuperscript{1} and Sugandhi.k\textsuperscript{2}
\textsuperscript{1}M.Tech .Vlsi Design, SRM University, Chennai, India
\textsuperscript{2}Asst.Professor (Sr.G), Department of Electronics and Communication SRM University, Chennai, India.

ABSTRACT

Big Data is very drastically every day growing the firms and other data pools. Importance of Big Data is enhancing the power of productivity, technology, analysis, design, business intelligence and Data Mining. In Big Data having drawbacks are capture, storage, search, sharing, analysis, visualization, less accurate patterns, less resolution, less performance and less quality of the clustering results. The proposed system uses this emerged the necessity of Big Data techniques as Hadoop. The cluster problems such as cluster characteristic or behavior, divide into parts, cluster validity or rationality, and cluster performance or presentation, scalability. Finally to evaluating the accuracy and quality of Hadoop clustering and MapReduce withBig Data set is Geo-logical data in finding the water resources and weather reports.

Keywords: Big Datasets, Hadoop, Clustering, Distributed Clustering, Data preprocessing, MapReduce, Big Data Technique.

1. INTRODUCTION

1.1 Introduction

Data generated in different organization, industries and other data pools. It is all about better Analytic on a broader spectrum classify something in terms of data, and therefore capability and possible to create even more differentiation among industry peers. Big Data facts are every day creates 2.5 quintillion bytes of data. Big Data Similar to ‘Small-Data’ but bigger data requires techniques, tools, architecture, Big
Data is more than just a DataWarehouse that requires to store and analysis large volume of data. The volume Big Data is the term for a collection of data sets so large and complex that it becomes difficult to process using on hand database management tools or traditional data processing applications. The challenges include captureing or gathering,curation or activity or process, data base maintained, search, sharing, transfer, analysis, and visualization. The trend methods and technology to large amount of datasets is due to the additional information derivable from analysis of a single large set of relevant fallowing 1. Big Data is about Massive data volume 2. Big Data run on Hadoop 3. Big Data means unstructured and structured data 4. Big Data is for social media feeds and sentiment analysis5. NoSQL means No SQL.Big Data is the opportunity to extract insight from an immense volume, variety and velocity of data.

Hadoop is a software framework, which means it includes a number of components that were specifically designed to solve large scale distributed data storage and maintained, analysis of performance and retried tasks. The Hadoop components are necessary for a Big Data solution of problems to solve and some of these components can be replaced with other technologies that better complement a user's needs to technology on Hadoop. The example is Hadoop distribution, MapReduce which includes NFS as another solution to HDFS offers a fully change any time access the read/write file database system. The Distributed process frameworks data base on MapReduce and scalable open source implementations of various analyses, clustering algorithms, to possible now to extract the knowledge and intelligence of Big Data sets.

Apache Software Foundation (ASF) to introducthat Apache Mahout, a new open source project with the primary goal of creating scalable clustering algorithms which can be use free under the Apache License. Mahout contains effective implementations for clustering, categorization, Mahout uses the Apache Hadoop library to scale effectively in the Mahout
Project was introduced as a sub project by several people involved in the Apache (open source) community which has active interest in clustering.

The cluster center residing in one map may have points within its preview that is part of the input data to another mapping function. The results from the Reducing are then finding into the appropriate Mapper and Reducer to begin the next round of processing. Hadoop Mahout is that implements developments some of the clustering and classification algorithms which have been modified to fit the MapReduce model. The Mahout implementations have been deployed within Apache Hadoop a Map-Reduce based cloud runtime. Mahout has been designed to work specifically with Hadoop, there is nothing to preclude using the Mahout library within data pre-processing database system to supports the MapReduce paradigm and clustering algorithms are an unsupervised machine learning technique that facilitates the creation of clusters, which allow us to group similar and relevant items (also called observations) together so that these clusters are similar in some definition of Big Data. Clustering has more broad large applications in part of Big Data areas and industries such as Data Mining, Database system, guidance of recommendation systems, design pattern recognition, identification of understanding clustering algorithms have certain unique characteristics in the Hadoop environment.
2. SYSTEM ANALYSIS

As data volumes increase the quickly becomes untenable to perform this clustering over a single machine. The challenge in implementing distributed clustering algorithms is that it is possible that an algorithm will get stuck in local optima, never finding the optimal solution. Attempting to converge on an optimal solution can be even more difficult when data is distributed, where no single node is fully aware of all data point.

Big Data process is to extract the data from web. the process of navigating on web pages in the domain of internet, to read and explore to care maintained data. Computing ready to use to follow a list of configure domains at steps of depth and particular number of pages. An order of the extracting files helps to searching the massive content in blazing fast. Fetching the data it has to mine and vector is so as to plot large amount of text. The basic techniques work out on the Data so as to satisfy the User Experience. The Raw data which is in unsupervised learning turn to be the useful data clustering data base maintained the user.

Figure 2.1: Proposed Architecture for Analysing Big Data Sets

2.1. Big Data

Big Data it is all about better analytic on a broader spectrum of data, Big Data Is more than just a DW that requires to store/analysis large volume of data. Big Data is not just about Volume of data that resides in DW today. The volume could be batch and real time on single definition “Big data is the term for a collection of data sets so large and complex that it
becomes difficult to process using on hand database management tools or traditional data processing applications.”

Figure 2.2: Big Data Challenges

The challenges are collecting, gathering, data importance of power, database maintains to searching, sharing, moving, analytics, and visualization. The new business technology, important of remember the quality to search, public, crime, social, determine the rules and regulations. Big Data comes from every day to day create 2.5 quintillion bytes of data 90% of the data in the world today has been created in the last two years data comes from anywhere sensors to gather climate information and

posts and share the to the social network medias, photos and videos and Purchase transaction records to maintained, cell phone GPS signals to record database administration.

Figure 2.3: Data Sources of Big Data

The Myths Big Data:

Big Data is new, Big Data is about large Data Volumes, Big Data means a Hadoop technique, and Big Data need a Data Warehouse, Big Data is for social networks and publics’ database maintained.
The Hadoop Modules

![Diagram of Hadoop Modules](image)

**Figure 2.4: Hadoop Eco-System**

The Hadoop Eco-system mainly there are Hadoop common, Hadoop Distributed File System (HDFS), MapReduce, Hadoop YARN, Hive, Pig, HBase, Storm, Avro, Mahout, Zookeeper.

The Hadoop module process through outputs on data framework for job scheduling and cluster resource on data management. Database system for parallel process of Big Datanumber of wayto supports data storage on data warehouse infrastructure that provides expiation and querying. A scalable machine learning and data mining dataflow language and execution design framework for parallel maintained the large performance relative service for different applications.

2.2. Apache Mahout

Apache Mahout is an open source library from machine learning on Apache Foundation. Mahout is to provide needed framework utility of usage for different programming. Mahout can be capable the huge amount of data compared to Clustering and Classification of machine learning. Apache Hadoop, Mahout can distribute and manage it’s commutation over a clustering of servers on dataset. Apache Mahout began Life in 2008 Sub Project for Apache Lucene project to first text mining, Mahout is a Library which supports machine learning and it takes some of the distribution and different technique in Hadoop so as to maintain Big Data.

![Diagram of Mahout in Hadoop](image)

**Figure 2.5: Mahout in Hadoop**
Mahout currently implements three areas that are quite commonly used in real applications:

- Collaborative filtering
- Clustering
- Categorization

Figure 2.6: Architecture of Apache Mahout

Log-likelihood Coefficient:

The next step is to identify the target neighbors. This is calculated using the threshold-based selection. In the threshold-based selection, items whose similarity exceeds a certain threshold are considered as neighbors of the target item. Then the final step is the prediction, to calculate the weighted average of neighbor’s ratings, weighted by their similarity to the target item. The rating of the target user $u$ to the target item $t$ is as following

$$P_{ut} = \frac{\sum_{i=1}^{c} P_{ui} \times \sin(t,i)}{\sum_{i=1}^{c} \sin(t,i)}$$

$P_{ut}$ is the rating of the target user $u$ to the neighbor item $t$, $\sin(t,i)$ is the similarity of the target item $t$ and the neighbor item $i$, and $c$ is the side of cluster. The forecast ranking is sorted and stores them in recommended list. The user ID and its corresponding recommend list as the middle key/value of output of the reduce phase. If the Hadoop platform has not enough location of Mapper handle, the platform has to hold ob Mapper complete its task and given its data, and then analysis of a new Mapper to work with the user documents file. The method to continue the all tasks are completely evaluated by the user.
3. SYSTEM DESIGN AND IMPLEMENTATION

3.1. Requirements

It deals with both the hardware and software requirements for the project. The requirements are given below.

Hardware Requirements

These include the basic hardware specifications needed for the system to run the application.

- **RAM**: 4 GB
- **HARDDISK**: 100 GB
- **Processor**: Core i3

Software Requirements

These include the software essential for running the project including the Operating System. Programming Language etc. for this project we require the following software.

- **O.S**: Linux Platform (Ubuntu 14)
- **Software’s**: VMware work station 10.6

3.1. Modules

The Big Data to performance of improvement of data clustering algorithms to use different modules they are Data pre-processing, MapReduce and clustering. Mahout’s is build scalable machine learning libraries. The algorithms for clustering, classification and batch based collaborative filtering are implemented on top of Apache Hadoop using the map/reduce paradigm contributions that run on a single node or on a non Hadoop cluster are welcome as well. The core libraries are highly optimized to allow for good performance also for non distributed algorithms. Mahout is distributed under a commercially friendly Apache Software license, scalable community.

3.1.1. Clustering

**k-Means Clustering**

The k-Means is a simple algorithm for grouping objects on database, clustering. All objects need to be represented as a set of all features. The user particular the
several of groups to identify feature vector in k points in that vector space, particular centres of the clusters. All objects are each assigned to the center of cluster a new center is computed by average the present vectors of all objects of data, the work of assigning objects and recomposed centers is repeated process converges. The algorithm can finite numbers of iterations, distance measure, initial center computation average centers have been explore, to estimation of the number of clusters k.

The script outline that will get you started with k-Means.

- Accepts clustering type: k-Means, fuzzyk-Means, lda, or streaming k-Means
- Gets the Geo-Logocal dataset
- Runs org.apache.lucene.benchmark.utils.Extractfile to generate out from geological.xml (the downloaded archive)
- Runs seqdirectory to convert Geological- out to SequenceFile format
- Runs seq2sparse to convert SequenceFiles to sparse vector format
- Runs k-Means with 40 clusters
- Runs clusterdumpshow on results

**Implementation**

The implementation accepts input directories one for the data points and one for the initial clusters. The data directory number of input files of SequenceFile(Key, VectorWritable), clusters more SequenceFiles(Text, Cluster) maintainer k initial clusters the input directories are modified by the development, the input data points are stored in 'testdata' and the 'output/clusters-0' directory. Driver executes running the k-MeansDriver the output directory directories: 'clusters-N'' contain the clusters for each iteration and 'clusteredPoints' clustered data points.
Figure 4.4: k-Means Algorithm

Invocation using Java involves supplying the following arguments:

**input:** a file path string to a directory containing the input data set a SequenceFile(WritableComparable, VectorWritable).

**clusters:** a file path string to a directory the initial clusters, a SequenceFile(key, Cluster \ Canopy). k-Means clusters and Canopy canopies may be used for the initial clusters.

**output:** a file path string an empty directory is used for all output from the algorithm.

**distanceMeasure:** Qualified class name of an instance of DistanceMeasure used for the clustering.

**convergenceDelta:** The Algorithm has converged (clusters have not moved more than the value in the last iteration)

**maxIter:** The Max number of steps to run, independent of the convergence specified

**runClustering:** A Boolean indicating, if true, clustering step is to be executed after clusters have been remember.

**runSequential:** The Boolean the k-means sequential implementation is to be used to
process the input data.

After running the algorithm, the output directory to contain are clusters-N.SequenceFiles(Text, Cluster) process by the algorithm every step. The Text key is a cluster identifier string clusteredPoints are (if clustering enabled) a directory SequenceFile(IntWritable, WeightedVectorWritable). The IntWritable key is the clusterId. The WeightedVectorWritable value is a double weight and a VectorWritable vector where the weight indicates the probability that the vector is a member of the cluster. k-Means clustering, the weights are computed as 1/(1+distance) the distance the cluster center and the vector using the chosen DistanceMeasure.
4. TESTING

4.1. Software Testing

The testing process is successfully executing program to find the software problems. Testing is conducting executes it cover errors to bugs represent in the software. Software testing is an essential of software quality maintains and assurance represents the number of ultimate reusable of the specification and identification of test results of program. Design and development of code testing and software reliability, comparability to during new definition implement stage it was attempted to build software from an a introduce concept to tangible measurable implementation of testing.

The aim of the testing process is to evolution identify all defects on software product. It is not possible to guarantee that the software occurrence given the assurance because of the fact that the input data domain the most of software projects of reducing defects errors in a system and increasing the user’s confidence and quality of product, given the assurance in a developed system on the project. Testing a process of subject the program to set of test inputs and observe the program to execute fails to as expected, then the conditions under problems, errors, failure of program occurs to identify after debugging and correction of testing.

Testing Objective and Test Principles

Testing is a process of executing the program to finding an error. A good test case is one that has a probability of identifying as a yet another defaults false and errors. The principals are Tests should be identifying to end user requirements. Tests should be planning to long time before testing starts. Testing should start on a little scale and results towards testing on the system.

Testing Strategies

The testing to set of activities that can be planning in advanced and conducted systematically strategies for software testing must be accommodate low level test that are necessary to verify that a small
source code segment has been correctly implemented as well as high level test the validation of major system function against customers requirement.

The Strategy for software testing integrates software test cases into a process of good planning steps to the result in the successful implements of software. Software testing is a behavior of topic of the program referred to as Verification and Validation. Verification refers to the set of activities of the software correct to implements step by step evolution of particular function and Validation refers the set of activities of the software that has been developments to be identifying traceable to user’s requirements.

### 4.1. Test Cases and Test Results

**Table 4.1: Test Case - Data pre-processing on Dataset**

<table>
<thead>
<tr>
<th>Test Case ID</th>
<th>TC_01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Case Name</td>
<td>Data Pre-processing</td>
</tr>
<tr>
<td>Test Case Description</td>
<td>Elimination of false data on dataset</td>
</tr>
<tr>
<td>Requirements Required</td>
<td>Weka tool</td>
</tr>
<tr>
<td>Expected Results</td>
<td>Given the format data set</td>
</tr>
<tr>
<td>Actual Results</td>
<td>Correct output is obtained</td>
</tr>
<tr>
<td>Test Result</td>
<td>Pass</td>
</tr>
<tr>
<td>Remarks</td>
<td>Nill</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Case ID</th>
<th>TC_02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Case Name</td>
<td></td>
</tr>
<tr>
<td>Test Case Description</td>
<td>Take input as a dataset</td>
</tr>
<tr>
<td>Requirements Required</td>
<td>Hadoop</td>
</tr>
<tr>
<td>Expected Results</td>
<td>Data set.</td>
</tr>
<tr>
<td>Actual Results</td>
<td>Correct output is obtained.</td>
</tr>
<tr>
<td>Test Result</td>
<td>Pass</td>
</tr>
<tr>
<td>Remarks</td>
<td>Nill</td>
</tr>
</tbody>
</table>

**Table 4.2: Test Case - Take Input as a Dataset**
5. RESULTS

The quality and accuracy of clustering algorithms on BigData sets using the Hadoop doing the first download the geological data sets do the data pre-processing to select the WEKA tool. In the WEKA tool installation after show the one pane are window. To select the dataset doing the data pre-processing on database, in the process of pre-processing to elements the duplicate values and noise data, massive values and default values deleted after comes the data set to storage. The storage dataset doing the MapReduce on dataset on Hadoop environment, after perform the Hadoop clustering the geological dataset. The dataset doing mahout various clustering, the dataset to evolution of execution times on clustering support vectors, centric of clusters, process in Avg, Min, Max, SD execution time of cluster, coefficient factor of clusters on dataset, to improving the quality on clustering on Hadoop environment based on windows platform comparison to improve the clustering results on Big Data sets on using Hadoop. The results are to finding the water resources and different areas and to show the weather report on system on dataset.

Figure 5.1: Data pre-processing for Geo-Logical Dataset

The WEKA tool select pre-processing on data set, what are the data set to select doing the Pre-processing. Todisplays the false and correct the data.
the clustering to select rows are columns on data set.

Figure 5.2: Data pre-processing Output
The selected data set select the row are columns are text data set on pre-processing doing to show the output of the pre-processing

Figure 5.3: Clustering on Dataset
The WEKA tool doing clustering on data set to show the results, in

Figure 5.4: Hadoop Task Tracker Status
The Hadoop to installation to show the configure the Hadoop task tracker

Figure 5.5: MapReduce on input File with Hadoop
The MapReduce on the file during mapping and reducing the system are input file configuration on processing on Big Data.

*Figure 5.6: Hadoop MapReduce Output Location*

To shows that the Hadoop MapReduce on data sets location of directory of system, it shows that mapping the file on type, size, replication, time show on MapReduce location on data set.

*Figure 5.7: Graphical representation of the comparison of all Clustering Algorithms*

The figure shows that graphical representation of comparison of all clustering Algorithms in based on the cluster in dataset on database to evolution of execution time and support vector of cluster in compression of different type of clustering.

*Figure 5.9: Improve the Performance of various Clustering Algorithms*

The figure shows that to improve the clustering on the Hadoop environment to comparisons of different algorithms performance evolution the canopy clustering given the best performance evolution of Hadoop clustering.
6. CONCLUSION AND FUTURE WORK

Clustering is designed and applying it overcome many clustering issues such as cluster tendency, cluster partition and cluster validity. Although many clustering algorithms are in Mahout and WEKA are unwanted elements to effective ones. In the mahout it performs clustering algorithms and evaluating, and also clusters cohesiveness and cluster quality of all the clustering algorithms in Hadoop Mahout by giving the regular Big Data sets. To evaluate the datasets in data pre-processing and Hadoop clusters and Hadoop map reducing on dataset to calculate the accurate the results and decries the evaluation time of data sets. The clustering algorithms on Mahout, now they are refined to effectiveness evaluated the performance, cluster cohesive and good quality results of all the clustering algorithms in Hadoop by giving the standard Big Data sets. In the given data set execution of data set to evaluate first Data pre-processing to eliminates the faults data, after perform the MapReduce and clustering the Hadoop environment. To find out the water resource identification different areas and show the weather report identify on areas. The future work is to develop storage on Hive or HBase to manage database and user query retrieval inauthentication of cloud.

REFERENCES


Yun Lu, Ming Zhao, Lixi Wang and Naphtali Rishe, “v-TerraFly: large scale distributed spatial data visualization with autonomic resource

