

# Mining of Massive Dataset

## Project 2 Report

Hao Jiteng, Zhou Lizhi, Yang Fangzhou

June 19, 2012

### Abstract

K-means is a simple yet useful clustering algorithm. It's underlying natural implies that it could be parallelized or distributized. Some implementation of kmeans employs platforms such as Hadoop and CUDA to boost the process of mining of massive dataset. In our project we implement k-means algorithm on Apache Hadoop Project. We ran our algorithm on our tiny cluster. Evaluation has been done to measure our algorithm.

## 1 Introduction

### 1.1 Hadoop

Following is the introduction of Hadoop on its project homepage [1].

The Apache<sup>TM</sup> Hadoop<sup>TM</sup> project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

We mainly use the Hadoop MapReduce framework rather than HDFS.

### 1.2 k-means

There are many materials that introduce k-means algorithms. MacQueen, J. et al. proposed k-means algorithm in [4]. In [3] the author introduced a simple k-means MapReduce algorithm.

The dataset is relatively large comparing to the memory size. Thus we need a mechanism to deal with the incompatibility of memory. MapReduce

is a solution to this problem. The detail of our algorithm is documented in the following section.

## 1.3 Dataset

### 1.3.1 MSRA image data set

The MSRA image data set contains 295 classes of images. Every image class has hundreds of images. 7 features are used to describe the images. There are files that describe the relevance between the image and the topic.

For our situation, we combine the 7 features into one and regards it as a vector with about a thousand dimensions. We didn't use the relevance information for simplicity. The evaluation results are shown in Section. ??

### 1.3.2 A-set dataset

The A-set dataset [5] is used to test our algorithm's correctness. This dataset is downloaded from the Internet. The visualization of this data set shows that the data points are relatively separated, which is very suitable to verify the correctness of our k-means.

## 2 Clustering Algorithm

In this section we discuss the k-means clustering algorithm used in our implementation. Note that to find proper initial clusters, which may lead to fewer iterations and good result, we also implemented a Canopy clustering MapReduce algorithm.

### 2.1 Canopy Clustering

Canopy method is a relative easy but efficient clustering method. The basic idea of Canopy is to put the similar items into the subset, which are called Canopies. Different Canopies can overlap each other, but there is no node, which doesn't belongs to any of Canopy. Thus, here we use this method to pretreat the dataset before Kmeans method.

The algorithm is built up by the following MapReduce phases.

#### 2.1.1 Procedure

1. Mapper

`<LongWritable, Text> → <1, CanopyCluster>`

Choose two distance threshold  $t_1$  and  $t_2$ , where  $t_1 > t_2$ . For each node P in the list, calculate the distance between node P and Canopy, if this distance is less than  $t_1$ , put P into this Canopy cluster, if no such kind Canopy cluster exists, take P as a new Canopy. If P was near

enough to some Canopy cluster, which means the distance between P and the Canopy is less than  $t_2$ , delete node P from the list.

## 2. Reduce

`<1, CanopyCluster> → <identifier, KmeansCluster>`

There is only one reducer and the reducer's job is to do another canopy clustering on the output of the mappers described above. That's why the mappers always assign 1 to the keys of every output pair. For this round of clustering we can use another  $t_1, t_2$  combination, namely  $t_3$  and  $t_4$ . It's handy to use the KmeansCluster as the result format and kmeans clustering pass could load the result quickly.

## 2.2 k-means Clustering

Mahout Project [2] is a data mining framework under Apache<sup>TM</sup> Foundation. It contains a k-means implementation. The blog [6] gives a very detailed view of the algorithm. Our algorithm mainly based on the idea of Mahout Project. Actually, this algorithm is very similar to the BFR algorithm in our lectures.

The algorithm is built up by the following MapReduce phases,

1. The k-means iteration, output every cluster centroid if converged.
2. Assign every point to a known cluster and output the result.

The detail of the first phase is described as follows

### 2.2.1 Procedure

#### 1. Mapper

`<LongWritable, KmeansCluster> → <LongWritable, KmeansCluster>`.

It reads the input file content as value. The key is the value offset in the file. Then it convert the value to a VectorDoubleWritable, which is used to represent the feature vector. It finds the nearest cluster to the vector, and output cluster id as key, a new cluster containing only point as value.

#### 2. Combiner

`<LongWritable, KmeansCluster> → <LongWritable, KmeansCluster>`.

It reads the output from Mapper and combine those tuples who have the same cluster id(meaning that they are assigned to the same cluster) using KmeansCluster.omitCluster(). This function reduced the network transmission flow because the actual meaningful information need to be communicated between different nodes are only the N, SUM and SUMSQ of clusters, which is described in the lecture slides of BFR algorithm.

### 3. Reducer

`<LongWritable, KmeansCluster>→<LongWritable, KmeansCluster>`.

It reads the cluster id as key and the KmeansCluster as value. It adds the N, SUM and SUMSQ. The result leads to the combination of clusters. Finally the reducer outputs the result clusters of this single iteration. This clusters are input of next iteration. During reducer, if the movement of one cluster is less than a threshold, it's said to be "Converged". If a cluster is converged, a counter in context will increase by one.

4. If, in the driver, the counter equals the number of clusters, meaning that all clusters are converged, this phase is finished.

#### 2.2.2 Assign Point to Clusters

After the iterations, the clusters are stable. Next we are going to assign every point to its nearest cluster. This phase only requires mapper to do all the works, since this procedure is highly parallelized. Each two points are independent of each other.

The mapper takes the following form:

`<LongWritable, Text>→<Text, Text>`

It takes the data offset in file as key and the data as value. First it converts the Text object into a VectorDoubleWritable object that represents a data point. For each point we find the nearest cluster and output the cluster id and some other useful information for statistic and evaluation as value.

## 3 Evaluation

To evaluate the results of our experiment, we calculate the precision and recall of the results, then calculate the F value. Better results we get will lead to a better score.

### 3.1 Evaluation Dataset

Since the whole MS image dataset is too large, we pick a subset of the files but each item contains all features. Here is the file list:

Also, we use a set of different convergence thresholds to evaluate the results and the running time.

### 3.2 Cluster Information

To run our Hadoop, we established a tiny cluster with 3 computers connected by a 100Mbps TP-Link router using Ethernet.

1. Namenode/DataNode/TaskTracker
  - Intel Core 2 Duo P8700 @ 2.53GHz
  - 2 Core 2 Threads
  - 4G DDR3 1066 RAM
  - OpenSUSE 12.1 x64
2. DataNode/TaskTracker
  - Intel Core 2 Duo T6600 @ 2.2GHz
  - 2 Core 2 Threads
  - 2G DDR2 800 RAM
  - Archlinux i386
3. DataNode/TaskTracker
  - AMD Processor @ 3.0GHz
  - 6 Core 6 Threads
  - 8G DDR3 1066 RAM
  - Ubuntu 12.04 LTS x64

### 3.3 Evaluation Results

#### 3.3.1 MSRA image dataset

Results are listed in the table.1

| THRESHOLD | PRECISION | RECALL | F | ITERATION | TIME |
|-----------|-----------|--------|---|-----------|------|
|-----------|-----------|--------|---|-----------|------|

Table 1: Evaluation Results

#### 3.4 Note on the Results

The results of given dataset are really poor, but we guess the dataset can be a problem because the features are actually not relarant to the labels.

We have run some testings on some different datasets whose points can be distinguish clearly. Our program gave almost perfect results.

##### 3.4.1 A-set result

As we can see in the dataset download webpage [5], our second evaluation dataset is quite seperated. It doesn't matter whether using canopy clustering, the kmeans clustering converge at once with a totally correct clustering result. It proved that our algorithm and MapReduce clusters are correct, no matter how bad the result of MSRA image dataset is.

## 4 Conclusion

We finished this project on a own built tiny cluster infrastructure with our own PC and laptops. We implemented a k-means algorihtm using hadoop framework with MapReduce. We tested our algorithm on our infrastructure with MSRA image dataset and 1024 dimension A-set dataset from [5]. The evaluation result shows that our algrihth converge fast although there are some inaccuraccy when using the MSRA data set.

## References

- [1] Apache Fundation. Hadoop project. <http://hadoop.apache.org/>.
- [2] Apache Fundation. Mahout project. <http://mahout.apache.org/>.
- [3] Ricky Ho. K-means clustering in map reduce. <http://horicky.blogspot.com/2011/04/k-means-clustering-in-map-reduce.html>.
- [4] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.
- [5] Speech and University of Eastern Finland Image Processing Unit, School of Computing. A-sets. <http://cs.joensuu.fi/sipu/datasets/>.
- [6] Leo Zhang. Learning mahout: K-means clustering. <http://www.cnblogs.com/vivounicorn/archive/2011/10/08/2201986.html>.