Achieving Load Balancing of HDFS Clusters Using Markov Model

Jin Kyu Kim
Soonho Kong

Problem

In matrix form, it is defined as follows:

Distance Function: Use \( m \) "jaccard distance" with \( k \) of files

Cluster Method: complete, average, single, weighted

Hierarchical Clustering

CMU OpenCloud - 64 Nodes

To Achieve Load Balancing. We need to Predict the files which will be accessed near Future.

Goal

220K Files Reused (33%)

Preprocessing Raw Data

Markov Model

Seq. of 10831 Hourly Logs

Encode using clusters

Transition Matrix \( A \)

Build Markov Model

Prediction

Markov predicts 99% of 220K files when we pick out 25% of total files for migration.

Experiments: Impact of Parameters

Experiments: Performance Comparisons

Markov's prediction precision improves up to \( x4 \) in comparison to Random.

Conclusion

The best clustering method is complete.