Avinash Pathrol

STUDENT, MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

Massive Unstructured Text Data Processing with Map Reduce Algorithms

The magnitude of data generated and shared by businesses, public administrations, and scientific research has increased immeasurably. On average every day the world produces around 2.5 quintillion bytes of data, with 90% of these data generated in the world being unstructured. Regardless of where it is generated from and shared, with the reality of huge data comes the challenge of analyzing it in a way that brings Big Value.

In this research, the Map-Reduce algorithm is used for the computation of massive amount of data. The MapReduce first splits the input files and starts up many copies of the program on a cluster of machines. The master is the special copy of the program and the rest are workers assigned by the master. Master picks idle workers and assigns each one a map function or a reduced function. The map workers read the content of the corresponding input and parse key-value pairs. Immediate key-value pairs of the map function are buffered in memory. Periodically the buffered pairs are written to the local disk partition into regions by a partitioning function. Reduce worker reads the buffer data from the local disk of the map workers and when the reduce worker has read all intermediate data from its partitions, then intermediate keys are generated, and then the keys are grouped together based on their occurrences. The reduced work iterates over the sorted intermediate data and for each unique intermediate key encountered it passes the key and the corresponding set of these intermediate values to the user's reduced function. The output of this function is appended to a final output file.

This implementation will help manage and analyze the large data set with concurrent processing and thereby producing results faster and more conveniently. Experiments were conducted on COVID-19 fake news detection from Twitter data. In the future, we would like to implement proposed algorithms in different applications.

Research Advisor: Dr. Baidya Saha, Assistant Professor