MGI34-08 Room:301A Time:May 19 14:30-14:45

Distributed data mining from time-series satellite imagery using Hadoop -experimental study for modeling of GIMMS NDVI-

Kou Nishimae[1]*, Rie Honda[1]

[1]Kochi University

A large amount of time-series images have been stored in the data archive in the field of remote sensing of the Earth and planetary bodies. We examined Hadoop and MapReduce, a framework for distributed data system, aiming at construction of a large-scale spatio-temporal data mining system in the Earth and planetary science filed,

In the experiments, we used GIMMS (Global Inventory Modeling and Mapping Studies) that is normalized difference vegetation index (NDVI) provided as a time-series imagery available for 25 years spanning from 1981 to 2006. The data set is derived from imagery obtained from the Advanced Very High Resolution Radiometer (AVHRR) instrument onboard the NOAA satellite series.

We examined two major processes of spatio-temporal data mining with this data: (1) extraction of time series data from time-series images and (2) temporal modeling using logistic function via Maximum likelihood (ML) method. These processes were implemented on the distributed system using MapReduce on Hadoop system composed of one master machine and 50 client machines and examined their scalability. The method is basically is same with our previous study (Nishimae and Honda 2012), however the number of client machines is 10 times larger than the previous study and the experimental condition is selected more carefully. The efficiency of the distributed system was examined as the ratio of execution time for single machine with single core to that with multiple client machines.

Experimental results showed that the efficiency was increased almost linearly with the number of client machines and cores in the case of modeling using ML method, however in the case of extraction of time-series data, the efficiency was not increased after a number of client machines exceeded 30. Furthermore, the number of core does not affected to the efficiency of the system in the data extraction. However, the efficiency of the whole system is expected to increase with number of clients and cores because the execution time for modeling is two orders of magnitude larger than the execution time for extraction.

In addition, we have examined higher knowledge discovery process such as clustering by using Mahout that is a library for machine learning available on Hadoop and MapReduce. Preliminary results will be shown in the presentation.

Keywords: Hadoop, distributed, spatio-temporal, data mining, vegetation index