<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7499912>

A MapReduce-Based Nearest Neighbor Approach for Big-Data-Driven Traffic Flow Prediction

Abstract:

In big-data-driven traffic flow prediction systems, the robustness of prediction performance depends on accuracy and timeliness. This paper presents a new MapReduce-based nearest neighbor (NN) approach for traffic flow prediction using correlation analysis (TFPC) on a Hadoop platform. In particular, we develop a real-time prediction system including two key modules, i.e., offline distributed training (ODT) and online parallel prediction (OPP). Moreover, we build a parallel k-nearest neighbor optimization classifier, which incorporates correlation information among traffic flows into the classification process. Finally, we propose a novel prediction calculation method, combining the current data observed in OPP and the classification results obtained from large-scale historical data in ODT, to generate traffic flow prediction in real time. The empirical study on real-world traffic flow big data using the leave-one-out cross validation method shows that TFPC significantly outperforms four state-of-the-art prediction approaches, i.e., autoregressive integrated moving average, Naïve Bayes, multilayer perceptron neural

networks, and NN regression, in terms of accuracy, which can be improved 90.07% in the best case, with an average mean absolute percent error of 5.53%. In addition, it displays excellent speedup, scale up, and size up.

Summary:

The idea of this paper was to utilize offline distributed training and online parallel prediction by utilizing a MapReduce-based KNN model for traffic flow prediction using correlation analysis (TFPC). KNN was chosen for a few reason. KNN is a very simple model, no many parameters need to be optimized (risk of overfitting) and it’s capable of processing very complex sets of data.

The model tried to analyze the correlation of space and time as an inherent feature of traffic flow in a complex urban transportation networks (e.g. Beijing). The assumption they made was that the traffic flow of the target road segment at the future time interval is closely related to that of the same road segment at the previous and current time intervals.

The accuracy of the method used in the model was over 90% in the best case, which significantly improved the efficiency and the scalability of traffic flow prediction.