K Nearest Neighbour Joins for Big Data on MapReduce: a Theoretical and Experimental Analysis

Given a point $p$ and a set of points $S$, the kNN operation finds the $k$ closest points to $p$ in $S$. It is a computational intensive task with a large range of applications such as knowledge discovery or data mining. However, as the volume and the dimension of data increase, only distributed approaches can perform such costly operation in a reasonable time. Recent works have focused on implementing efficient solutions using the MapReduce programming model because it is suitable for distributed large scale data processing. Although these works provide different solutions to the same problem, each one has particular constraints and properties. In this paper, we compare the different existing approaches for computing kNN on MapReduce, first theoretically, and then by performing an extensive experimental evaluation. To be able to compare solutions, we identify three generic steps for kNN computation on MapReduce: data pre-processing, data partitioning and computation. We then analyze each step from load balancing, accuracy and complexity aspects. Experiments in this paper use a variety of datasets, and analyze the impact of data volume, data dimension and the value of $k$ from many perspectives like time and space complexity, and accuracy. The experimental part brings new advantages and shortcomings that are discussed for each algorithm. To the best of our knowledge, this is the first paper that compares kNN computing methods on MapReduce both theoretically and experimentally with the same setting. Overall, this paper can be used as a guide to tackle kNN-based practical problems in the context of big data.