Chapter 7
Parallel Queries of Cluster-Based k Nearest Neighbor in MapReduce

Wei Yan
Liaoning University, China

ABSTRACT

Parallel queries of k Nearest Neighbor for massive spatial data are an important issue. The k nearest neighbor queries (kNN queries), designed to find k nearest neighbors from a dataset S for every point in another dataset R, is a useful tool widely adopted by many applications including knowledge discovery, data mining, and spatial databases. In cloud computing environments, MapReduce programming model is a well-accepted framework for data-intensive application over clusters of computers. This chapter proposes a parallel method of kNN queries based on clusters in MapReduce programming model. Firstly, this chapter proposes a partitioning method of spatial data using Voronoi diagram. Then, this chapter clusters the data point after partition using k-means method. Furthermore, this chapter proposes an efficient algorithm for processing kNN queries based on k-means clusters using MapReduce programming model. Finally, extensive experiments evaluate the efficiency of the proposed approach.

INTRODUCTION

The *k* nearest neighbor query (*k*NN query) is a classical problem that has been extensively studied, due to its many important applications, such as knowledge discovery, data mining, and spatial databases. With the rapid growth of the spatial data, parallel *k*NN queries are a challenging task. MapReduce programming model processes large scale datasets by exploiting the parallel and distributed computing parallelism. The MapReduce programming model provides good scalability, flexibility and fault tolerance. Therefore, MapReduce programming model becomes an ideal framework of processing *k*NN queries over massive spatial datasets. This chapter proposes a method of parallel *k*NN queries based on k-means clusters using MapReduce programming model.

DOI: 10.4018/978-1-4666-9834-5.ch007

The *k* nearest neighbor query (*k*NN) is a special type of query that is *k* nearest neighbors from points in *S* for each query point *r* in dataset *R*. The *k*NN query typically serves as a primitive operation and is widely used in knowledge discovery, pattern recognition, and spatial databases. Now, lots of researches (Yao *et al*. 2010) have been devoted to improve the performance of *k*NN query algorithms. However, all these approaches are performed on a single, centralized server. In single machine, the computational capability and storage are limited, and its efficiency is low. How to perform the *k*NN query on parallel machines is an important issue in cloud computing environments.

ALL the existing work has concentrated on the spatial databases based on the centralized paradigm. Xia *et al*. (2004) proposed a novel *k*NN-join algorithm, called the Gorder *k*NN join method. Gorder is a block nested loop join method that exploits sorting, join scheduling and distance computation filtering and reduction to reduce both I/O and CPU costs. It is simple and yet efficient, and handles high-dimensional data efficiently. However, the system of centralized server will eventually suffer from performance deterioration as the size of the dataset increases. A solution is to consider the parallel query processing in distributed cloud computing environment.

Parallel spatial query processing has been studied in parallel database, cluster systems as well as cloud computing platform. In cloud computing environments, a large part of data-processing using MapReduce (Dean *et al*. 2004) programming model runs extensively on Hadoop. The MapReduce programming model provides a powerful parallel and distributed computing paradigm. Cui *et al*. (2014) addressed the problems of processing large-scale data using *k*-means clustering algorithm and proposed a novel processing model in MapReduce to eliminate the iteration dependence and obtain high performance. A data structure that is extremely efficient in exploring a local neighborhood in a geometric space is Voronoi diagram (Okabe *et al*. 2000). Given a set of points, a general Voronoi diagram uniquely partitions the space into disjoint regions. The region corresponding to a point *p* covers the points in space that are closer to *p* than to any other point.

This chapter presents a partitioning method using Voronoi diagram that is multi-dimensional spatial datasets partition into Voronoi cell. This chapter uses *k*-means clusters method to apply on the Voronoi cell. The *k*-means algorithm is a well-known method for partitioning n points that lie in the *d*-dimensional space into k clusters. Then, this chapter proposes a method of pivot for the Voronoi diagram-based data partitioning, which uses the *k*-means clusters algorithm to choose as pivot. Furthermore, this chapter proposes an efficient algorithm for processing *k*NN queries based on *k*-means clusters method using MapReduce programming model.

The objectives of the chapter are summarized as follows:

- This chapter proposes a partitioning method using Voronoi diagram that is multi-dimensional spatial datasets partition into Voronoi cell.
- This chapter proposes a *k*-means clusters method to apply on the Voronoi cell. The center point for each cluster is chosen as pivots of Voronoi diagram.
- This chapter proposes an efficient algorithm for processing *k*NN queries based on *k*-means clusters method using MapReduce programming model.