International Journal of Decision Support System Technology
Volume 16 • Issue 1

A Novel Query Method for Spatial Database Based on Improved K-Nearest Neighbor Algorithm

Huili Xia, College of Computer and Artificial Intelligence, Zhengzhou University of Economics and Business, China*
Feng Xue, College of Computer and Artificial Intelligence, Zhengzhou University of Economics and Business, China

ABSTRACT

Spatial database is a spatial information database and is the core component of geographic information systems (GIS). Aiming at the problem that time complexity of k-nearest neighbor (kNN) querying algorithms are proportionate to scale of training samples, an efficient query method for spatial database based on the Spark framework and the reversed k-nearest neighbor (RkNN) is proposed. Firstly, based on the Spark framework, a two-layer indexing structure based on grid and Voronoi diagram is constructed, and an efficient filtering and a refining processing algorithm are proposed. Secondly, the filtering step of proposed algorithm is used to obtain the candidates, and the refining step is used to remove the candidates. Finally, the candidate sets from different regions are merged to get the final result. Results of experiments on real-world datasets validate that the proposed method has better query performance and better stability and significantly improves the processing speed.

KEYWORDS

Big Data, Geographic Information System, Parallel Processing, Reverse K-Nearest Neighbor, Spark, Spatial Database

INTRODUCTION

A spatial database is a database system that describes, stores, and processes spatial data and the associated attribute data in which a relational database management system (RDBMS) is used to regulate spatial data, mainly solving the data interface problem between the spatial data stored in the relational database and the application program—that is, the spatial database engine (SDE) (Sveen, 2019; Baharin, & Akunne, 2020). More precisely, spatial database technology is used to address the accessing issue of the geometric attributes of the spatial data objects in the relational database (Breunig et al., 2020; Zhang et al., 2022). In addition, the spatial database can also effectively handle the integrating, querying, and managing of complex spatial information for huge volumes of data. Compared with traditional databases, spatial databases have wider

DOI: 10.4018/IJDSST.332773 *Corresponding Author

application prospects and are the core components of a geographic information system (GIS) (Choi et al., 2020; Xia et al., 2022).

With the continuous development of location-based services (LBS), spatial databases are of great significance in the fields of traffic network systems, GIS, and various decision support systems (Pant et al., 2018). The query technology of spatial data is the utmost important operation for a spatial database. Spatial data query technology refers to the process of searching for and returning of spatial objects that meet certain query constraints in a spatial database. Commonly used spatial database query operations include Exact Match Query (Jiang et al., 2019), Point Query (Wang et al., 2019), Region Query (Wan et al., 2019), and Nearest Neighbor Query (Chen et al., 2019).

The problem of nearest neighbor query is a hotspot in today's database technology, and the query results are usually spatial data or datasets that meet the query requirements. Nearest neighbor query is the basis of the spatial database query field. Because of the continuous change of query requirements, many variants have been produced, such as k-nearest neighbor (kNN), directional nearest, clustered nearest, group nearest, constrained nearest, and privacy-preserving nearest neighbors (Wang et al., 2019). The kNN and reverse nearest neighbor (RkNN) algorithms are considered to be the most fundamental and widely used query types. The kNN algorithm can be provided as an LBS service alone, such as finding the nearest restaurant or hotel, or it can be used as a basic technology to provide support for other queries, such as RkNN and search window query (Xu et al., 2021). RkNN has a good application when it comes to the "influence range" and "influence degree" of an object on other objects, such as the location problem when building new infrastructure, such as shopping malls and hospitals, and the current popular concept of targeted advertising.

Most of the current parallel RkNN query algorithms are on the basis of the Map-Reduce framework. Ji et al. (2015) introduced a distributed RkNN query processing algorithm on the basis of the inverted grid index. García et al. (2019) conducted distributed RkNN query research on Spatial Hadoop, the spatial expansion framework of Hadoop, and proposed an RkNN query algorithm MRSLICE based on Spatial Hadoop. MapReduce framework along with open-sourced applications in Hadoop are breakthroughs in the efficient processing of large-scale datasets in which data is processed through a distributed system, and program execution will not be affected by node failures. However, researchers have found in experiments that the performance of multicore devices is limited by the single node implementation, which can easily lead to too tight coupling and low scalability. The limitation of MapReduce in the Hadoop platform will cause the start-up overhead of each round of operation in the iterative calculation process to be too large, and the overhead of disk I/O will also reduce the execution efficiency (Grolinger et al., 2014). In contrast, the Spark framework overcomes the aforementioned problems, and it accelerates the processing of distributed computing through memory computing strategies. This framework allows users to store data in memory for repeated queries, making it more practical for online, iterative, and dataflow algorithms (Meng et al., 2016).

In recent years, scholars have proposed frameworks such as SpatialSpark (You et al., 2015) and GeoSpark (Yu et al., 2015). These frameworks can realize Spark-based distributed spatial range query, kNN query, and spatial join query, and it is verified by experiments that Spark-based query processing is better than the MapReduce framework. García et al. (2017) presented a query algorithm on the basis of Spatial Hadoop (RkNN-SH) and implemented it on actual datasets. In addition to the above typical spatial queries, scholars have expanded the research on variant queries based on the Spark framework, including distance join queries (García-García et al., 2020), spatio-temporal join queries (Li et al., 2021), and top-k spatial join queries (Qiao et al., 2020).

The above studies have demonstrated the superiority of the Spark framework in processing parallel spatial queries. However, more research on RkNN query based on Spark framework needs to be done. To this end, we propose an efficient query method for spatial databases based on the improved RkNN algorithm. This method includes the following innovations: