Research on the identification of channel congestion based on big data of navigation

Shijie Yan

Navigation College, Shandong Jiaotong University, Weihai, China

Keywords: AIS system; nautical data; channel congestion; DBSCAN; parallelization

Abstract: The development of big data has always been the current and future direction of development. The emergence of large data technology has immediately attracted the attention of the navigation industry. It hopes to use large data technology to solve the problems in the mining of navigation information and to better master and use navigation data resources. It is in the background of this research that this topic comes into being. Based on a thorough understanding of the connotation of navigation data, this paper analyzes the problem of channel congestion in depth, explores the use of navigation data to solve the problem of channel congestion identification, and puts forward a channel congestion identification method of Spapa erkk-DSBCA. The feasibility and effectiveness of this method for channel congestion identification are verified by experiments.

1. Introduction

Large data refers to huge data over the conventional storage level. Thanks to the promotion of the Internet of things in the field of intelligent transportation and navigation information, the automatic acquisition of navigation data has been based on some new technologies, such as intelligent channel, intelligent port, ship network, ship traffic service (VTS) and other new sources. A large amount of data has been generated, and maritime related government departments and enterprises have accumulated a lot of navigation information resources. However, due to the lack of a unified data platform, a large number of data are scattered and scattered, and a large number of valuable data have to be deleted regularly, and the value of data can not be fully exploited by the lack of data mining methods. Therefore, as the "big data" technology appeared, it immediately aroused the attention of the navigation industry, and hoped to solve the problems of navigation information fusion, storage, processing, inquiry and analysis with the help of large data technology, so that the large data of navigation played an important role in the decision-making of the government and enterprises.

The development of nautical large data is not only limited to the data we have mastered, but rather a resource that seeks to extract the most valuable information from the data, and turns "a pile of data" into "large data" in the most feasible way. The development of large data has become an irreversible trend. It is of great significance to make navigation related governments and enterprises grasp the development opportunities brought by the big data, to master and use the data resources better and to excavate the information of commercial value.

This paper studies how to conduct real-time monitoring of the fairway and maximize the avoidance of waterway congestion so as to enable a large number of ships to navigate smoothly through the channel. On the basis of the large data of navigation, this paper thoroughly analyzes the traffic flow conditions such as the speed, flow and traffic density collected by AIS and GPS equipment, and the real-time analysis and control of the traffic flow according to the monitoring information of the channel, so that the traffic flow and the ship can be accurately mastered. The state of navigation is timely released according to data analysis. We should avoid the congestion of the fairway and give full play to the navigation function, improve the efficiency of navigation and reduce the traffic accidents caused by congestion.

2. Description of channel congestion

The large data of navigation refers to the mass data produced in the fields of navigation, management, supervision and so on, and the general name of the related technologies and solutions about the effective fusion, storage, processing, query and analysis around the scale of the data. Shipping is big data, because it also has the characteristics of "4V":

1) The amount of data is large. Nautical data is constantly being generated worldwide, and data volume will jump from TB to PB and EB level.

2) There are many types of data. Because of the diversity of navigation data, the data can be divided into structured data and unstructured data. The text data and multimedia data in nautical data belong to unstructured data.

3) The value density is low. There is an inverse relationship between the value density of data and the total amount of data.

4) The processing speed is fast. Navigation data are often collected automatically from the first line of production. Due to the needs of production and transaction, the transmission, storage and processing of data are required to be completed in a short period of time, especially in port supervision. When it is found that illegal behavior needs to be stopped immediately, it is very demanding for the handling capacity of the large number of seafaring.

The great value of nautical data embodies great academic appeal, which can be transformed into a rich set of information to help people to learn, develop and maintain the ocean.

In recent years, due to the rapid development of shipping industry, the variety of ships has been developing to a variety, large and high speed, which has also promoted the increase of the harbour facilities, but it also brought about the problems related to the congestion of the channel. There are many ways to control the traffic congestion, but the most important thing is how to get the information of the management area accurately and promptly, and then take the corresponding measures to adjust it. Therefore, real-time acquisition of information in channel congestion area is very important for marine traffic guidance and control. The channel plays an important role in the water transportation economy. With the increase of the water transport volume, the problem of the congestion of the channel gradually becomes prominent, which has become the concern of the channel management. The factors that affect the passage capacity are divided into natural factors, channel factors, shipping factors and management factors.

1) The influence of natural factors on channel passing capacity is mainly influenced by the number of normal navigable days in the channel. The natural factors mainly include meteorological factors and hydrological factors, including wind, fog, wave, water, tide and so on.

2) The channel factors mainly refer to the influence of waterway depth, channel width, channel length, navigation rules, navigation and navigation aids on the normal navigation of the channel.

3) Ship factors mainly refer to the ship's ton level and average loading coefficient, ship navigation speed, ship density based on the ship field, the imbalance of ship to port and the influence of its ship.

4) the management factors mainly refer to the management of ship accident, the way of shipping in and out of the channel, the information management of the navigable condition in the channel, and the management of the port dispatch, which are related to the conditions of the channel, the level of the traffic management in the channel and the driver factors.

In this paper, the clustering analysis algorithm is used to cluster the ship's ship data in the navigation data, and the ship cluster formed by the ship clustering in the navigation channel is obtained in real time, and the traffic flow velocity is used to judge the traffic jam section. Clustering data of ship data comes from AIS system.

3. Large data Spark computing platform

The arrival of the era of big data has also led to the constant updating of the big data processing platform ecosystem. With the birth of HadoopHDFS, HadoopMapReduce, HBase and Hive, the

early Hadoop ecological circle was gradually formed. Hadoop, as an open source project for Apache, aims to build a stable, scalable and distributed large data processing platform using cheap hardware.

First of all, we understand the shortcomings of other distributed systems (such as MapReduce), which are mainly manifested in duplication, combination, limitations of scope of application, uneven distribution of resources and management. Thus, a new class MapReduce computing framework Spark was born. Spark is a memory based computation, and its processing speed is much better than that of MapReduce in many iteration calculations. As a general large data computing platform, Spark has become an integrated and diversified large data platform based on the concept of "OneStacktorulethemall". It is easy to deal with real-time flow computing, SQL interactive query and machine learning in large data processing platform. Practice and graph calculation.

Spark was officially open source in 2010, became the Apache fund project in 2013, and became the top project of the Apache fund in 2014. The whole process was less than 5 years. Spark is the first three of all open source projects in the Apache foundation. Compared to other large data platforms, the Spark code base is most active, as shown in Figure 1.

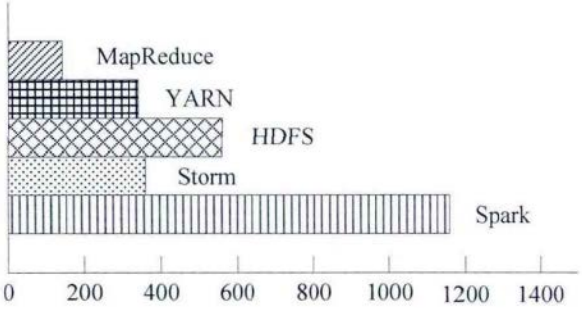

Figure 1 Code activity of a large data computing platform

At present, the development of Spark big data computing platform has generated many sub projects. Berkeley University calls the entire ecosystem of Spark the Berkeley data analysis stack (BDAS), as shown in Figure 2 below. On the basis of the core framework Spark, it mainly provides four categories of computing frameworks and other computing frameworks, including SparkStreaming, Graphx, MLbase, SparkSQL, etc., of which, SparkStreaming is the main branch. Graphx is a parallel graph computing framework; MLbase mainly supports the underlying distributed machine learning and machine learning functions; SparkSQL is a consulting engine for the SQL query and analysis supporting structured data mining. It is because of these top sub projects that Spark has provided a more high-level and richer computing mode.

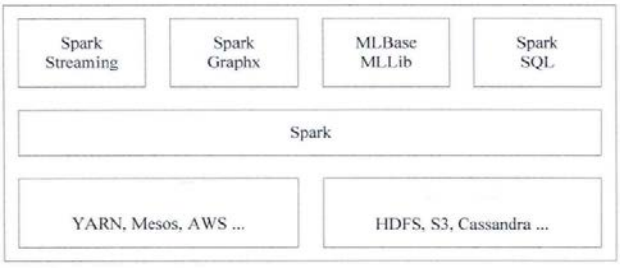

Figure 2 Spark data technology stack

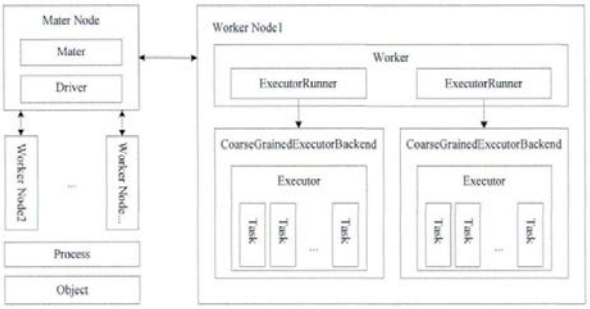

Figure 3 The deployment diagram of Spark

To understand a new system architecture, one of the first things to do is the deployment of the system. After the successful deployment of the system, each node will start the whole system related services. What services can be started by each node can be seen from the figure 3. In the entire Spark cluster, it is divided into Master nodes and Worker nodes, which correspond to Master and Slave nodes in Hadoop, which are Hadoop.

4. Research on DBSCAN algorithm based on Spark framework

DBSCAN algorithm is used to solve the problem of clustering process. The application process is to divide the set of data objects processed into different threshold values according to the different methods of measuring the similarity threshold, and the similarity threshold of the data object in the same threshold domain belongs to the same threshold. The range of values. On the contrary, the similarity between the data objects in different threshold ranges is lower, that is, the higher the similarity degree is adjusted in the same cluster domain by cluster domain classification, the lower the similarity adjustment between different cluster domains is, the better the calculation results of the clustering process.

The first DBSCAN algorithm is proposed by EsterMartin et al. The process of density clustering is the process of extracting the threshold space with high similarity degree in the low similarity threshold space. In the calculation process, the density of the data object in the region is identified by the threshold density and the distance measure to determine whether the data object is in the threshold area. The distance density is based on the distance of the central node to calculate the data density. In the DBSCAN algorithm, the method of calculating the density of a specific data object is to judge the statistics according to the data density of the radius region of the object itself.

The traditional DBSCAN algorithm ignores the volume of a single data object and regards each data object as a particle, but this method does not apply to the problem of ship traffic clustering. Because there are all kinds of ships in the sea channel, the size of the ship is usually different, and the size of the ship is related to the degree of congestion. Therefore, the shape of the ship is simplified to rectangle.

In order to improve the efficiency of DBSCAN algorithm, this paper uses parallel processing method, extends the DBSCAN algorithm to the Spark platform and parallelization. That is, it is processed separately on multiple processors, and then the individual clustering results are merged.

Every moment in the world, there are one hundred thousand kinds of ships sailing on the four oceans, and more than 5000 ports around the world, generating huge amounts of information. This experiment is to apply the design of this paper to the AIS data set in the world to calculate the congestion of the channel in real time. The algorithm is applied to actual channel congestion analysis, and good results are obtained.

References

[1] Liu Yu Xiao. Recognition of channel congestion based on fuzzy comprehensive evaluation [D]. Zhejiang Gongshang University, 2014

[2] Li Hao. Development and research of shipping information platform based on the Yangtze River

trunk line AIS. Wuhan: Wuhan University of Technology, 201L. D.

[3] Liang Yan. Research on Parallelization of data mining algorithms based on distributed platform Spark and YARN [D]. Zhongshan University, 2014

[4] Gao Yanjie.Spark big data processing: technology, application and performance optimization [M]. Machinery Industry Press, 2015.

[5] Tang Zhenkun. Design and implementation of machine learning platform based on Spark [D]. Xiamen University, 2014

[6] Wei Jinrui. A class of model-based clustering methods [J]. Statistics and information forums, 2014, 29 (2): 19-22