ATLANTIS PRESS

Advances in Computer Science Research, volume 91

International Conference on Modeling, Simulation and Big Data Analysis (MSBDA 2019)

Construction and Application of Ship Data Mining Platform Based on Spark

Lei Cao, Jingfeng Hu* and Ran Li

Dalian Maritime University, Dalian 116026, China

*Corresponding author

Keywords: Spark, data mining, Automatic Identification System, platform construction, Data preprocessing.

Abstract. With the explosive growth of ship information data brought by the intellectualization and digitization of waterborne traffic, it has become more and more important to process and apply massive ship data quickly and accurately. In view of the growing maturity and wide application of Spark large data technology in data mining field, this paper, against the characteristics of ship AIS data, chooses a series of large data technologies based on Spark to build the AIS data mining platform. According to the basic process of mining and processing, the overall framework of the platform is divided into three modules: database, Spark computing and mining, visualization. The AIS data of Sunda Strait with high ship density is used to process and mine the actual data to verify the availability of the platform. The results show that the platform works well and the three modules can work normally. And the AIS data of the Strait are processed and mined quickly and accurately. According to different requirements, it can display bar charts, polyline charts, histogram, trajectory charts and thermal charts and other commonly used data analysis graphics to achieve the effect of navigation assistance.

Introduction

Nowadays, the world economy is a globalized economy, and sea, land and air transportation has been greatly developed. Because of its advantages of heavy load, low cost and wide cross-region, water transportation occupies 90% of the transport market and occupies an absolute dominant position [1]. With the development of intellectualized and digital technology in shipping industry, blowout data volume comes along. The traditional data processing platform has been overburdened. The emerging big data technology brings us dawn. For dealing with the increasing data, the big data technology has unique advantages, no matter its rapidity, accuracy and real-time, which can not be compared before. Therefore, in order to strengthen navigation monitoring and regional flow analysis, this paper proposes to build a Spark-based ship data mining platform to process and mine the surging ship data. Based on database module, mining calculation module and visualization module, the platform realizes rapid and efficient mining of AIS data from a series of operations such as data input, pre-processing, storage, output, mining calculation and visualization, and generates a series of effect maps, which are helpful to the decision-making of platform users.

In order to verify the validity of this platform, ship data of Sunda Strait, one of the busiest waterways in the world, are selected for processing, excavation and verification. Sunda Strait is located between Sumatra Island and Java Island in Indonesian archipelago. It is about 120 kilometers long and 22-110 kilometers wide. Its average water depth is much higher than that of Malacca Strait, which is very suitable for large vessels [2]. It has promoted the global trade circulation and made great contribution to the interchange of various countries and regions, which has also made the Sunda Strait one of the most important traffic detection points in the world [3].

Construction of AIS Data Mining Platform

Platform Environment Configuration and Software and Hardware Introduction

(1) Environmental Configuration

In this paper, the construction of big data mining platform chooses to install and run software such as spark and Hadoop in dual system. Running the spark in a dual-system environment, the test environment is more accurate, reliable, and avoids the conflict between the virtual machine IP address and the local IP address that may occur when the virtual machine is installed. It can be used for big data mining in the Linux environment. The environment configuration and construction of the platform will not conflict with the Windows system [4].

The large data mining platform studied and built in this paper is based on Ubuntu 14.04. Ubuntu has beautiful user interface, perfect package management system and rich technology community. Ubuntu also has good compatibility with most hardware, including graphics cards and so on.

(2) Introduction of Software and Hardware

Hardware section: Graphics workstation (equipped with Ubuntu operating System).

Software section: as shown in Table 1:

Table 1.Related softwares of AIS data mining platform

No.	Software	Function
1	HBase	Data storage
2	Hive	Data warehouse tools
3	HDFS	Distributed file system
4	SQL	Data query
5	Spark	Computing engine
6	Scala	Programming language
7	Carbondata	Column data organization
8	IDEA	Programming platforms
9	YARN	Resource manager
10	Hadoop	Bigdata platform

Composition of Platform Module

The construction of this platform mainly focuses on three modules: database module, Spark data mining module and visualization module. The AIS data received from ships on shore are connected to the data input interface, stored in the Hbase database, and the basic data preprocessing is carried out directly in the database, such as speed judgment, whether the location of latitude and longitude points is on land and so on. Spark data mining module calls the data that has been preliminarily pre-processed in the database, carries out the next processing and mining, mining different forms of returned data according to the needs, and then links the visualization module to display the corresponding effects in the visualization module, including polygons, bar charts, trajectory charts, thermal charts and other effect charts, for users to observe and make decisions. The module running process is shown in Figure 1.

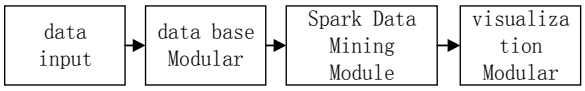

Figure 1. AIS data mining platform module flow chart

The flow chart of the database part of the large data mining platform built in this paper is shown in Figure 2. The data collected in real time are directly input into the Hbase database through the data flow interface and stored in the initial table A. Then, the basic pre-processing work is carried out directly in the database. For example, if the ship speed exceeds 35kn, it is judged as a data error; the data on the land is judged to be an error, and so on. The filtered data are stored in the next level table named Available Table M and Non-Available Table N. The data used in subsequent data mining and front-end display are clean data in Table M.

This module is the core module of this platform. It loads various data processing and mining algorithms in the module, and further processes and mines the AIS data that has been pre-processed in

the database for further data front-end display. This section will introduce the basic process of the algorithm already loaded on this platform. In the future development, the algorithm of the platform can be improved and added in the next step, and the scalability is very strong, which ensures the adaptability and generality of the platform. The flow chart of the data mining process is shown in Figure 3.

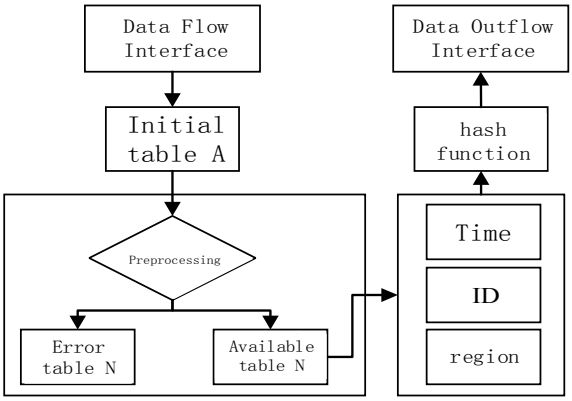

Figure 2. AIS data mining platform database module flow chart

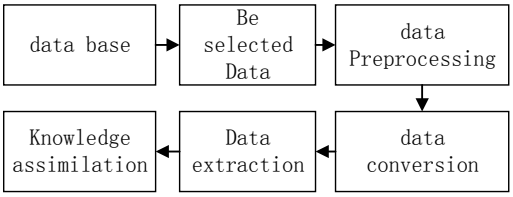

Figure 3. Flow chart of data mining

The visualization module of this platform is based on Echarts architecture, using JavaScript language for front-end display and debugging, compatible with most of the current browsers (IE8/9/10/11, Chrome, Firefox, Safari, etc.). The bottom layer relies on the lightweight vector graphics library ZRender, which provides intuitive, interactive, highly customizable data visualization charts, and can run smoothly on PC and mobile devices [5].

Echarts provides a regular line chart, histogram, scatter plot, pie chart, K Diagrams, box diagrams for statistics, maps for geo-data visualization, thermal attempts, diagrams, diagrams for relational data visualization, TreeMap, Sunrise diagrams, parallel coordinates of multidimensional data visualization, and BI Funnel Chart, dashboard, and support for mashups between diagrams and diagrams [6]. In addition to the built-in charts that contain rich features, Echarts Custom Series is also available. ECharts also provides rich graphical examples and an active developer community to meet the visualization needs of the vast majority of users.

Data Processing and Mining of Traffic Flow in Sunda Strait

The data selected in this paper are traffic flow data in Sunda Strait area from January to June, 2018. In this paper, M4 algorithm, Ifouce algorithm and ant colony algorithm are used to process the data of ships passing through the Strait, and the information of ship type, tonnage and country of ship origin are deeply excavated. The corresponding visual information such as pie chart, stack chart and thermodynamic chart are obtained.

Analysis of Ship Types and Tonnage Passing through Straits

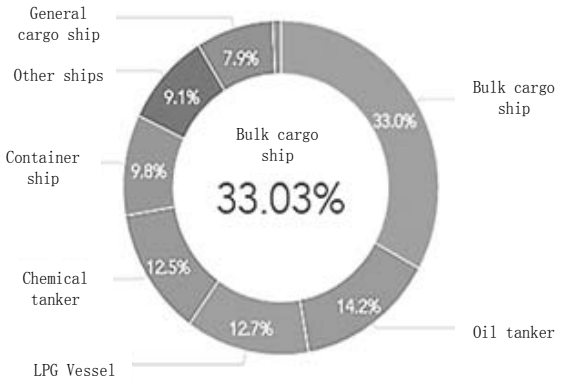

Figure 4. Total relation pie chart of ship type and tonnage

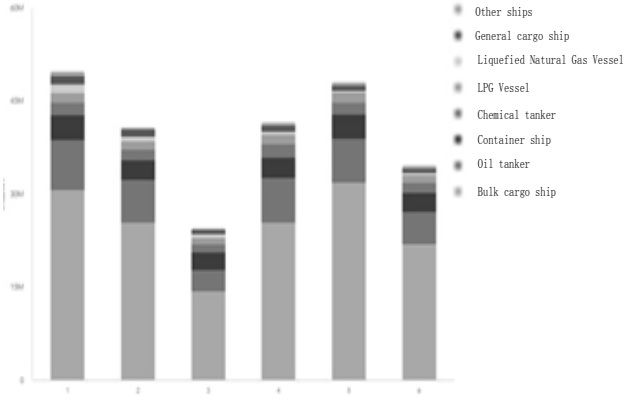

Figure 5. Stacking column chart for ship type and tonnage

Figure 4 shows the relationship between the type of ship passing through the Strait and the total tonnage, and the tonnage of the area through the region within the set time dimension is approximately 2.4 billion tons, tonnage percentage of bulk cargo ship accounted for the vast majority, reached the 62.6% , followed by tankers 15.9% , container ships 8.3% , chemical vessels 4.4% , liquefied petroleum gas vessels 3.6% , liquefied natural gas vessels 1.6% , miscellaneous cargo ships 2.32% and other ships 1.2%.

Figure 5 for the time bar diagram of ship type tonnage through straits and the relationship between ship type and tonnage total, bulk cargo accounts for the vast majority of each month's tonnage, followed by tankers, container ships, chemical vessels, liquefied petroleum gas vessels, liquefied natural gas vessels, grocery ships and other ships.

Analysis of Origin Country, Type and Tonnage of Ships Passing through the Straits

Figure 6. Tonnage thermodynamic map of the country of origin

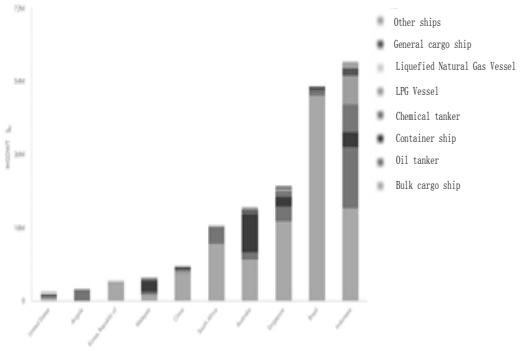

Figure 7. Accumulation column of ship type and tonnage of the country of origin top10

Figure 6 shows the thermodynamic maps of the countries of origin of all ships passing through the Straits within the set time dimension. As can be seen from the figure, in the source country, tonnage exceeds 1000kt mainly located in Southeast Asia, Australia, South America, China, Canada, southern Africa and other countries and regions, tonnage in 100kt to 1000kt mainly located in northern Africa, southern North America, the Middle East and other regions . By statistics, all countries through the region of the ship belong to a total of the data level of most tonnage totals in thousand and million.

Figure 7 shows the relationship between ship types and the top 10 tonnage of all ships passing through the Straits in the given time dimension. The sum of its tonnage exceeds the total of the regional tonnage 86.70%. Among them, Indonesia accounts for 24.63% , Brazil accounts for 22.12% , Singapore accounts for 11.89% , Australia accounts for 9.69% , South Africa accounts for 7.89% , China accounts for 3.56% , Malaysia accounts for 2.41% , Korea accounts for 2.19% , Angola accounts for 1.29% , US accounts for 1.03%.

Analysis of Destination Country, Type and Tonnage of Ships Passing through the Straits

Figure 8. Tonnage thermodynamic map of the country of destination

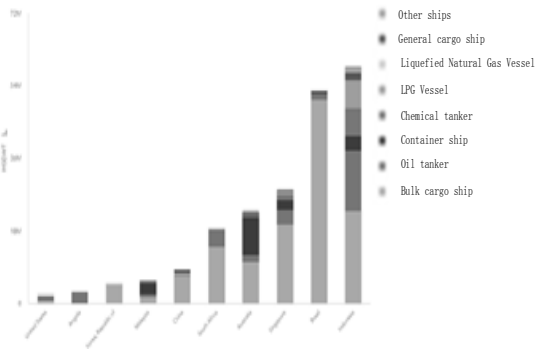

Figure 9. Accumulation column of ship type and tonnage of the country of destination top10

Figure 8 shows the thermodynamic maps of the destination countries of all ships passing through the Straits within the time dimension set. As can be seen from the figure, in the destination country, tonnage exceeds 1000kt mainly located in Southeast Asia, Australia, South America, China, India, southern Africa and other countries and regions, tonnage in 100kt to 1000kt mainly located in the United States, Russia, northern Africa, the Middle East and other regions. All other places are 100kt below, while Canada, the eastern coast of Africa are largely outside the destination country.

Figure 9 shows the relationship between the type of ship and the top 10 tonnage of all ships passing through the Straits in the time dimension. The total tonnage exceeds 97.83% of the total tonnage passing through the Straits. Among them, China accounts for 30.00%, Indonesia accounts for 25.93%, Australia accounts for 10.61%, Singapore accounts for 9.33%, Korea accounts for 4.50%, Brazil accounts for 3.23%, South Africa accounts for 2.98%, Japan accounts for 2.35% and Malaysia accounts for 2.31%, Taiwan accounts for 1.67%.

Thermodynamic Diagram of Traffic Traffic Trajectory through Straits

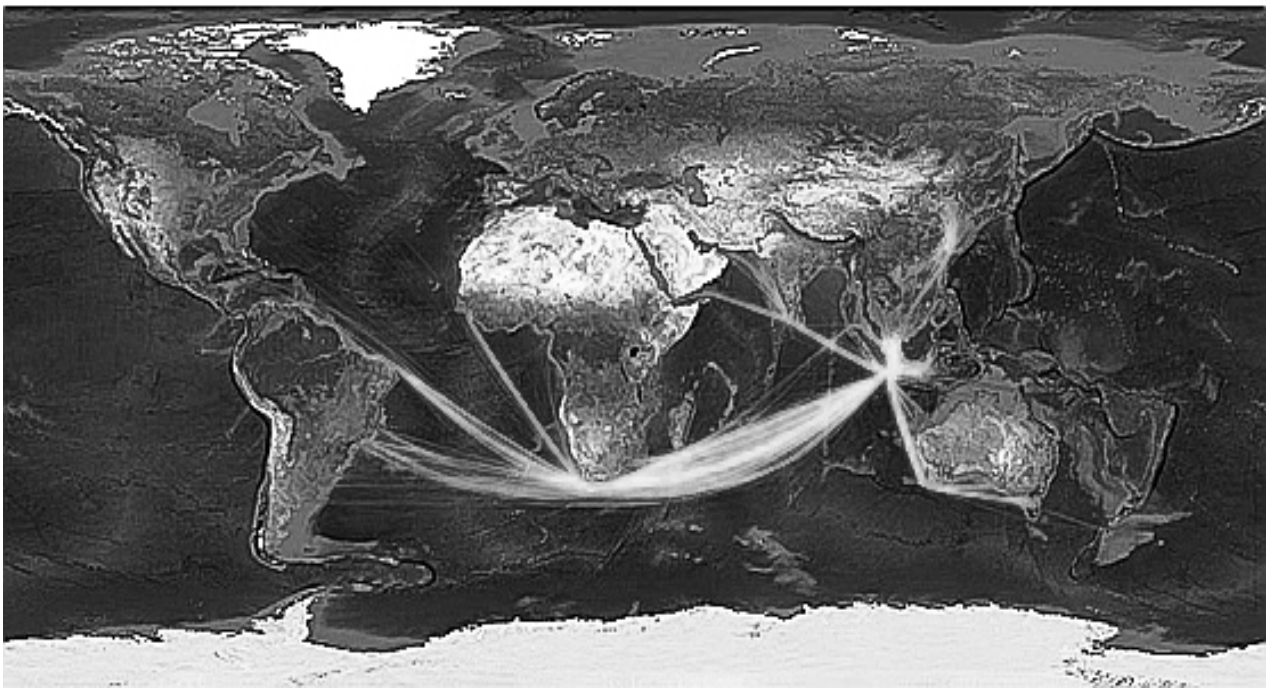

Figure 10. Thermodynamic chart of trajectory

Figure 10 shows the global trajectory through the Strait, in many countries of origin and destination, more concentrated in South America, Australia, Southeast Asia, China, India, the Middle East, Southern Africa and other regions, the star-shaped channel as the center of radiation to five continents the countries of origin and destination.More than 200 million tons of cargo and more than 5000 voyages pass through the Sunda Strait every year.Therefore, it is also particularly important to maintain the safety of navigation in the Sunda Strait, to ensure the safe movement of cargo ships and passenger ships, and to promote trade between the continents [7]. For China, as an important source and destination country of the Sunda Strait, as an important channel for China to deepen its ties with Southeast Asia, the Middle East, South America, North America, Africa and other countries, in order to promote the positive development of China's shipping industry, analysis and excavation of traffic flow data in the region, Ensuring the navigation safety of our ships in the Sunda Strait is a top priority.

Conclusion

In this paper, we build an AIS data mining platform and use the data of Sunda Strait to test and display the actual platform operation. The following results are obtained. (1) This paper chooses the Ubuntu operating environment under the dual system to design and build a large data mining platform for ships, and carries out a series of software installation and compatible debugging related to the

platform, including Spark, JDK, Hbase, Carbondata, etc. (2) The platform built in this paper is mainly divided into three modules: database module, Spark calculation module and visualization module. The main presentation modes of this platform are trajectory chart, thermal flow chart and other intuitive display trajectory, visualization of flow data; (3) Aiming at the AIS data of Sunda Strait, a series of automation operations such as data storage, pre-processing, data output screening, Spark mining calculation and front-end visualization are completed on this platform from the data access end, and the visualization effect of mining results is good.

Acknowledgement

This research was financially supported by the National Science Foundation of China (61772102), the Fundamental Research Funds for the Central Universities (3132016322).

References

[1] Shang Sinian, Design and implementation of massive AIS message data mining system based on cloud computing and distributed technology[D].Dalian Martime University,2017

[2] Xiang Zhe, Shi Chaojian, Hu Qinyou, Yang Chun. An approach for calculating competitive degree among ports using AIS data[J].Journal of Shanghai Martime University,2016,37(01):60-64.

[3] Zhu F, Maritime D . Mining ship spatial trajectory patterns from AIS database for maritime surveillance[C]// IEEE International Conference on Emergency Management & Management Sciences. IEEE, 2011.

[4] LIU Xiaobo, JIANG Yangsheng, TANG Youhua. Innovation platform of integrated transportation big data application technology[J].Big Data Research,2019, 4(6): 2018063.

[5] Handayani D O D , Sediono W , Shah A . Anomaly Detection in Vessel Tracking Using Support Vector Machines (SVMs)[C]// International Conference on Advanced Computer Science Applications & Technologies. IEEE, 2014.

[6] Kenthapadi K, Mironov I, Thakurta A G. Privacy-preserving Data Mining in Industry[C]//Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. ACM, 2019: 840-841.

[7] Stanisz T, Kwapień J, Drożdż S. Linguistic data mining with complex networks: a stylometric-oriented approach[J]. Information Sciences, 2019.

[8] Zabihi M, Pourghasemi H R, Motevalli A, et al. Gully Erosion Modeling Using GIS-Based Data Mining Techniques in Northern Iran: A Comparison Between Boosted Regression Tree and Multivariate Adaptive Regression Spline[M]//Natural Hazards GIS-Based Spatial Modeling Using Data Mining Techniques. Springer, Cham, 2019: 1-26.

[9] Gayen A, Pourghasemi H R. Spatial Modeling of Gully Erosion: A New Ensemble of CART and GLM Data-Mining Algorithms[M]//Spatial Modeling in GIS and R for Earth and Environmental Sciences. Elsevier, 2019: 653-669.