Fast data analysis method based on AIS system

Wenzhe Li[1,a],Kezhong Liu[2,b],Guo Jing[3,c]

[1]School of Navigation,Wuhan University of Technology,China

[2]School of Navigation,Wuhan University of Technology,China

[a]764042591@qq.com,[b]63327026@qq.com,[c]649616909@qq.com

Key words: AIS database, spatio-temporal data query

Abstract: With the development and perfection of AIS system, AIS data with high timeliness, strong stability and wide coverage provide a strong data base for traffic flow statistical analysis. This paper use Java with full consideration of JVM performance, to achieve rapid AIS data decode, then put forward a new method to improve the efficiency of spatial and temporal data query. The main contents of this paper as follows:

(1) This paper use reflection, multi-threaded mechanism to simplify the program structure to ensure thread independence and the atomic nature of the abnormal isolation, the use of random non-blocking IO storage technique to speed up data storage, reduce memory usage.

(2)extending Geohash and Base32 encoding applications, propose a method to compressed search space, thus improving the query efficiency.

Introduction

Automatic Identification System is an automatic continuous broadcasting system based on Very High Frequency.Based on the AIS system, ships quickly exchange all kinds of navigation data, so that maritime departments and research institutions have accumulated a lot of AIS data with high timeliness, stability and wide coverage. It provides a strong data base for maritime related research.Find and debug the existing analytical procedures, and ask the user experience and comments, found that the existing analytical procedures are mainly structural redundancy, low efficiency, large memory usage, unable to cope with the error or abnormal data and not timely updates and other issues, making it impossible to quickly and efficiently complete the massive message analysis or parsing task is no longer suitable for the current system under the task.

A fast method of AIS Message analysis

AIS Message analytical process

Message timestamp by password before 10 is given, its value as a starting time of the message to accept the long shore platform time (seconds), the starting time is 1970-01-01 00:00:00.The encapsulation information is intercepted according to the standard format of the AIS message, which is the parsing object.First, a binary stream is obtained by converting the package information to binary code by character one by one according to table 1.The first 6 bits of the binary stream are intercepted and converted to the decimal message ID, and the message type is judged according to the message ID. On the basis of the message type, cutting the binary data stream, for each field corresponding to the binary code and the field will be converted into decimal values, on the basis of the field and explained the construction

processing function, processing function field value converted to clear code, finally check the rationality of each clear code.

Table 1 Character conversion six bit binary code

Character conversion binary code pseudo-code
Input:character *ch* **(1)** *outSix*=ch+0x28**;** **(2)** if *outSix*>0x80 **(3)** *outSix*+=0x20**;** **(4)** else **(5)** *outSix*+=0x28**;** **(6)***outSix*=*outSix*<<2**;** **Output:**six bits binary code ***outSix*;**

Analysis result display and contrast test

Figure 1 is the main interface of the parsing program.

Figure 1 AIS original Bureau parser interface

EsiAisParser is a component based AIS parsing Java, this paper selected three size groups were 5M, 50M, 500M data for the analysis of efficiency of the testing procedures and EsiAisParser, analytical results show that the last time the analysis program is about EsiAisParser 1/10, the analytical efficiency 10 times.

Spatial data area query solution

Fast query of regional data based on Geohash

Neighborhood query based on Base32 coding.Geohash is an algorithm for encoding geographic data. Based on the idea of two points, it can quickly label spatial data by recursion. Tags can be used as index for spatial data, and general labels are stored in Base32 code[1].The query efficiency using the Geohash tag is far higher than the search for the latitude and longitude[2].

Table 2 is the pseudo algorithm of the Geohash algorithm. The algorithm inputs the longitude and latitude coordinate point and the reasonable binary stream length L, and the

output binary stream will be encoded by Base32, so L must be a multiple of five. According to the description of routines, it can be found that Geohash does not strictly and accurately map the position coordinates. The result of Base32 encoding represents a rectangular area. The location coordinates are located in the region, so Base32 encoding can be used as the label of the coordinate.Base32 coded prefix represents larger geographical area, such as encoding wx4g0ec1 and wx4g0, indicating larger area containing wx4g0ec1, namely neighborhood of wx4g0ec1[3].As Base32 code is introduced into the database, fuzzy query can be used to search a ship near a ship conveniently and efficiently.

Assuming that a ship is located in the wx4g0ec1 area, the following statement can be used to query the neighbourhood ships: the SELECT * FROM location reports -1 WHERE base32 LIKE'wx4g0e%'.Generally, when the Base32 code is 6 bits, the neighborhood is 1 kilometers near.

Table 2 Algrithm 1 Pseudo Code for Geohash

Pseudo code of Geohash algorithm
Step1: Define function HANDLER() to be called later ; **Input: *coordinate(X,Y)*;** **Output:"0" or "1";** **(1)**if *X*<*coordinate*<(*X*+*Y*)/2 return "0"; **(2)**else return "1"; **Step2: Start function GEOHASH():** **Input: the *coordinate* (*lon*,*lat*);** **legnth of the encoded binary string *N*** **Output: *binaries*** **(1)** define *m*=-180,*n*=180,*p*=-90,*q*=90,empty string *binaries* **(2)** while(*N*>**0**) **(3)** if HANDLER(*m*,*n*,*lon*)== "0" **(4)** append "0" to *binaries* **(5)** *n*=(*m*+*n*)/2 **(6)** else **(7)** append "1" to *binaries* **(8)** *m*=(*m*+*n*)/2 **(9)** if HANDLER(*p*,*q*,*la*t)== "0" **(10)** append "0" to *binaries* **(11)** *q*=(*p*+*q*)/2 **(12)** else **(13)** append "1" to *binaries* **(14)** *p*=(*p*+*q*)/2 **(15)** *N*=*N*-1 **(16)** end while

Spatial data area query method

Assuming that the irregular area of the user query is called Area, the key step of the regional data query is to identify all the Base32 codes associated with the Area.Table 2-12 gives a pseudo algorithm for a regional data query solution.In this, vertex_A is all the vertices of Area as the input of the algorithm.Each Base32 code represents the rectangle region Region. In step (2), INVERSE_GEOHASH is encoded by Base32 to get the coordinates of the four vertices of Region, and the process can be obtained by the geohash algorithm.Step (1) to (13) traverses the Region associated with each Base32 encoding search and Area in the database, and the correlation between the two is determined by (10) and (12).If Region is related to Area,

a simple query is established in Region, and the location report in Region is obtained, and the query result is saved to datas, and datas is used as output to give all location reports in Area.

Figure 2 gives 5 ways to relate Region and Area. In ①④, Region has a vertex in the Area.In ②③,Area has a vertex in the Region.In ⑤,Region and Area have edge intersecting.Step (10) determine whether Region is associated with Area in ①②③④.Step (12) determine whether or not the two are related in ⑤.

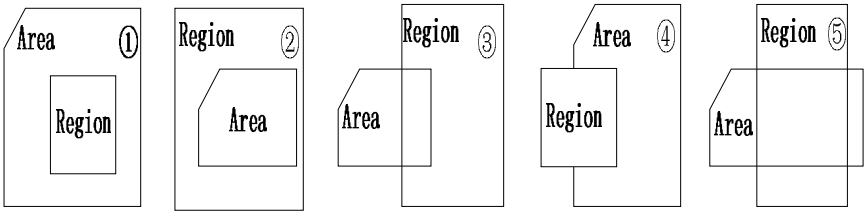

Figure 2Association method of Area and Region

According to the property of the vector product, if $\overrightarrow{P_0P_1} \times \overrightarrow{Q_0P_2} < 0$, then point Q_0 is clockwise on the line segment $\overrightarrow{P_1P_2}$.Assuming the vertices of a convex polygonP_1、P_2、P_3 and N points clockwise. If Q_0 satisfies: for any $i \in [1,N]$,there is $\overrightarrow{P_iP_{i+1}} \times \overrightarrow{P_iQ_0} > 0$.Then Q_0 are in the clockwise side of the convex polygon of each side, then the Q_0 in the interior of the convex polygon.This principle is based on the judgment of step (10), and the purpose is to judge whether Area and Region are in the other side. When the vertices of the two lines are respectively located on each side of the other side, the two lines must be intersected, $(\overrightarrow{P_1P_2} \times \overrightarrow{P_1Q_1}) * (\overrightarrow{P_1P_2} \times \overrightarrow{P_1Q_2}) > 0$ ensures that Q_1 、 Q_2 are located on $\overrightarrow{P_1P_2}$ both sides, $(\overrightarrow{Q_1Q_2} \times \overrightarrow{Q_1P_1}) * (\overrightarrow{Q_1Q_2} \times \overrightarrow{Q_1P_2}) > 0$ ensures thatP_1、P_2 are located on $\overrightarrow{Q_1Q_2}$ both sides.Step (12) determine whether the boundary between Area and Region is intersected accordingly.

Table 3pseudo code for regional data screening

Region data filtering pseudo code
Input: ***vertex_A*****;** **Output:list** ***datas*** **which contains datas in Area;** **(1)** for each Base32 code *b* in AISDATA **(2)** *vertex_R* = INVERSE_GEOHASH(*b*); **(3)** for *m*=1:length(*vertex_R*) **(4)** set *chosen*=true **(5)** Q_1=*vertex_R*[*m*] **(6)** Q_2=*vertex_R*[*m*+1] **(7)** for *n*=1:length(*vertex_A*) **(8)** P_1=*vertex_A*[*n*] **(9)** P_2=*vertex_A*[*n*+1] **(10)** if $\overrightarrow{P_1P_2} \times \overrightarrow{P_1Q_1} > 0 \;\&\&\; \overrightarrow{Q_1Q_2} \times \overrightarrow{Q_1P_1} > 0$ **(11)** if $(\overrightarrow{P_1P_2} \times \overrightarrow{P_1Q_1}) * (\overrightarrow{P_1P_2} \times \overrightarrow{P_1Q_2}) > 0 \;\|\; (\overrightarrow{Q_1Q_2} \times \overrightarrow{Q_1P_1}) * (\overrightarrow{Q_1Q_2} \times \overrightarrow{Q_1P_2}) > 0$ **(12)** *chosen*=false **(13)** if *chosen* **(14)** select as *RegionData* in AISDATA where base32=*b* **(15)** *data*=SELECTION(*vertex_A*,*RegionData*) **(16)** add *data* to *datas* **(17)** continue step(1) **(18)** else **(19)** continue step(3) **(20)** end for

Figure 2table 2-12 steps (10) and (12) judgment schematic

Determination of Region and Area related, step (15) according to the Base32 encoding the Region query all position report in Region data set RegionData, step (17) in SELECTION vertex_A Area according to the determined mathematical description, and then according to the description on RegionData data set query.In Figure 2-10 Xiazhimen channel as an example, the regional vertex (122.219902, 29.835830), (122.311882, 29.768110), (122.289541, 29.742870) and (122.210155, 29.829278), then describe the inequality in this area is:

$$\begin{cases} lat < -0.7362470 * lon + 119.8198674 \\ lat > 1.1297614 * lon - 108.4151361 \\ lat > -1.0884539 * lon + 162.8493958 \\ lat < 0.6722068 * lon - 52.3212232 \end{cases}$$

Figure 3 Xiazhimen channel

Assuming the screened Region encoding and Xiazhimen channel correlation for wx4g0ec1 and wx4g0ec2, the query statement:

```
select * from select * from (Location report -1
where base32 like ‘wx4g0ec1’ or base32 like ‘wx4g0ec2)
      where lat< –0.7362470 * lon +119.8198674
and Lat>1.1297614*Lon-108.4151361
and Lat>-1.0884539*Lon+162.8493958
and Lat<0.6722068*Lon-52.3212232;
```

query performance evaluation

In 2015 all position report -1 as the test data set, the query graph 5ten from July 28th to December 21st in the waters edge data query in 55 seconds, the query results of large amount of data, so to the right side of the heat map display. The same query is done again. The query does not use the database view, shields the base32 code, uses the conventional query scheme, and the whole query process takes at least 20 minutes. In contrast, the time and space data query scheme of this paper has increased the query speed by an order of magnitude. Analyze the specific implementation process. In fact, for more complex areas, the query efficiency of this scheme is higher, and the speed will increase with the increase of Base32 encoding length.

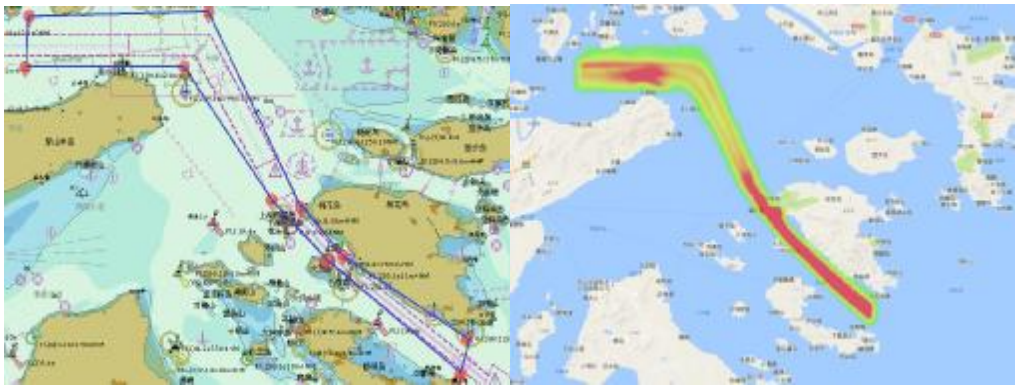

Figure 4 Data space query test area and results

Conclusions

In this chapter, a software system for data processing and visualization of AIS is designed and implemented. The basic architecture of the system is introduced. The main work is as follows:

(1) in view of the problem of low parsing efficiency and low fault tolerance in the current AIS parsing program, a data parsing scheme is designed to realize the fast parsing of the original AIS data.

(2) The Base32 coding of location data is constructed by Geohash, so that the database has the ability of fast neighborhood query. We further use the Base32 encoding to give the selection scheme of the region associated with the query area, search data in the related area, compress the search space, make the search process focus, and significantly improve the efficiency of spatial query. This scheme extends the application of Geohash and Base32 coding in spatial data query.

References

[1] Balkić Z, Šoštarić D, Horvat G. GeoHash and UUID identifier for Multi-Agent systems[M]. Springer Berlin Heidelberg, 2012: 290-298.

[2] Jin A, Cheng C Q, Song S H, et al. Regional query of area Data based on geohash[J]. Geography and Geo-Information Science, 2013, 29(5): 31-35.

[3] Jin A, Cheng C Q, Song S H, et al. Regional query of area Data based on geohash[J]. Geography and Geo-Information Science, 2013, 29(5): 31-35.