ISBN 978-981-11-7861-0
Proceedings of 2018 the 8th International Workshop on Computer Science and Engineering (WCSE 2018)
Bangkok, 28-30 June, 2018, pp. 141-145
doi: 10.18178/wcse.2018.06.025

A Grid-based Marine Traffic Hotspot Detection Algorithm on SpatialHadoop

Bao Lei [1 +], Yang Le [2]

[1] Computer Science Department, Wuhan East Lake College,Wuhan,China
[2] College of Electronics Engineering ,Navy University of Engineering,Wuhan, China

Abstract. Owe to the establishment of modern navigation and communication networks, maritime vessel trajectory data becomes increasingly available. These data sets are always huge and usually heavily skewed, finding hot spots area among maritime traffic data is critical for real time applications ranging from military surveillance to transportation management.In this paper, we proposed a grid-based hotspot detection algorithm on maritime traffic data. By gridding the spatial area into small buckets, it take the vessels density on each bucket into further calculation instead of the location of each single moving vessel. In order to handle the big data set, the computation apply the SpatialHadoop framework which can establish the R-tree spatial index to enhance location data handling performance on MapReduce programming. Experiments on real AIS data shows that the method propose is effective and fast.

Keywords: location big data, hotspot detection, trajectory data mining, SpatialHadoop

1. Introduction

Marine traffic hotspot refer to the area where vessels visit frequently.The purpose of hotspot detection is to find these hotspots out automatically by some methods,which can help the understanding of marine traffic situation in various applications such as traffic monitoring[1][2] and security surveillance[3]. The moving vessels' traffic data are large trajectory data sets which have large time span and across huge spatial area, the distribution are usually heavily skewed and the data always affected by various aspects such as hydro-geological environment, weather,crew and ship status.Related research mainly focuses on maritime trajectories data mining, but the few existing studies that examine the traffic hotspot were not sensitive to spatial characteristics,and even not designed to process big data.Aiming at this problem,this paper introduced a hotspot detection algorithm based on the MapReduce structure,which incorporate spatial index to handle spatial vessel location data. Experiment was done on the real Automatic Identification System(AIS) data and the result indicated that the method has effective detection result and high efficiency.

2. Related Works

The marine traffic data are very large location data sets, the data items on a single day can be millions,so the calculation on marine traffic data should use the big data frameworks such as Hadoop or Spark,but these traditional big data structures are ill equipped in supporting spatial data as it deals with spatial data in the same way as non-spatial data. The programs defined through map and reduce cannot access the constructed spatial index. To incorporate spatial data, some spatial big data model was proposed, including Hadoop-GIS[4],Spatial Spark[5] and SpatialHadoop[6]. Among them, SpatialHadoop is a comprehensive extension to Hadoop that injects spatial data awareness in each Hadoop layer. It adapts traditional spatial index structures to form a two-level spatial index for MapReduce environments, which has native support for spatial data available as free open-source.

+ Corresponding author.
E-mail address: blnj2000@163.com

The existing hotspot detection methods build statistical models on location data and use them on many applications such as crime site discovery, traffic surveillance. These methods works on relatively small data sets and can be divided into two categories: the point-based methods and the area-based methods.The point-based method use statistical methods or clustering models on moving points directly and find hotspot in them[7]. The area-based methods split the spatial space into basic area unit,further calculation use the features such as density,points number on each unit to build hotspot detection models[8].The point-based methods are more accurate but need far more computation because they need to take every single points into calculation. While the area-based methods are faster because they only need to sum up the features on each area unit, and further calculation only take the units into account.

In this paper, an area-based method to detect marine traffic hotspot is proposed, it split spatial space into small grid buckets, then use R-tree spatial index on SpatialHadoop to compute the traffic density on each bucket from large marine traffic location data, the hotspot regions are finally obtained from a cropping algorithm on these buckets. The method we present here is a fast algorithm which can incorporate large data sets and is fit for the spatial data.

3. Algorithm Preliminary

3.1. Available maritime traffic data

The location data used in this paper is from the Automatic Identification System(AIS).AIS is a tracking and self-reporting system used by maritime vessels to exchange information with other ships, AIS base stations, and satellites. AIS data consists of dynamic, static and voyage related information. Dynamic information, such as vessel heading, course, speed, position,etc., is broadcast in near real-time depending on vessels' speed and heading change. Static information (vessel identity,dimensions, etc.) is transmitted every 6 minutes as is voyage related data such as vessel destination, hazardous nature of its cargo, etc. In this paper, we use mainly the position data in out hotspot detection algorithm.

3.2. Hotspot in marine traffic

The whole maritime area $\boldsymbol{A}$ is gridded into $\boldsymbol{N=n*m}$ small buckets. The density $\boldsymbol{\rho_A}$ of the whole region $\boldsymbol{A}$ is the ratio of vessel numbers $\boldsymbol{Q_A}$ with the area of it $\boldsymbol{S_A}$.

$$\rho_A = \frac{Q_A}{S_A} \tag{1}$$

The hotspot $\boldsymbol{x}$ is an interested area which have a certain density of vessels activity, its detection is to find the region in the whole region $\boldsymbol{A}$, when the region $\boldsymbol{x}$ in $\boldsymbol{A}$ have a density $\boldsymbol{\rho_x}$,with:

$$\rho_x \geq h \times \rho_A \tag{2}$$

Then we call $\boldsymbol{x}$ a hotspot inside $\boldsymbol{A}$,and $\boldsymbol{h}$ is the hotspot threshold.

3.3. Spatial data indexing on SpatialHadoop

As the location data set is relatively large,to calculate the vessel density in each buckets,we need to build the spatial index and calculate it on SpatialHadoop framework, as shown in figure 1.

The spatial index we used is the R-tree spatial index,the establishment of R-tree index on SpatialHadoop need three steps: data splitting, location indexing and global indexing.The data splitting cut the initial data set into data slices according to there spatial neighborhood,and each slice has the size needed for HDFS storage.The local indexing build R-tree index on each data slice,and it can be used in fast data retrieval on local data slice. The global index is used to find the corresponding local data node,and is stored in the main node of SpatialHadoop.

When calculating the vessel density on a certain region, according to its location,the data retrieval process wont take the irrelevant data slices calculation by means of R-tree index we established.This will greatly increase the efficiency of our algorithm.And then the relevant data is parallel processed in the MapReduce framework : the number of map function is corresponded with the buckets number when we gridding the whole area, and each map function got a reduce function,in which the statistical calculation on the vessel density is done.

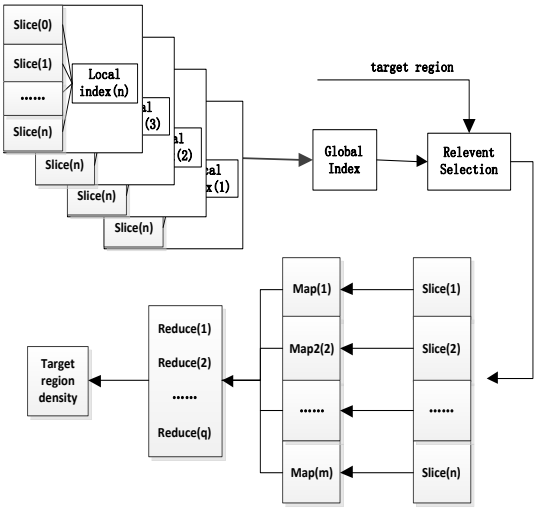

Fig. 1: The spatial indexing and retrieval of maritime traffic data

ALGORITHM 1 THE HOTSPOT DETECTION ALGORITHM

Algorithm1：Grid_HostspotDetection(*A*,*h*,*G*)

Input： ***A***:Target Area, ***h***:hotspot threshold,***G***:gridding size

Output ： ***S*** hotspot region in ***A***

1： Build R-tree index on whole region ***A*;**

2： Split ***A*** equally into ***G*=*n**m** grid buckets alongside the longitude and latitude;

3： Calculate the vessel number ρ_{ij} in each buckets on SpatialHadoop;

4： Let ***k***=0;
Let Initial candidate hotspot $H_k=A$;
Compute the density of candidate hotspot ρ_k

5： For all row i and column j
cut row A_i or column A_j of buckets off from H_k,
The result is hotspot region H_{k+1},
which will have the highest density ρ_{k+1},
if multiple rows or column result in the same density,
cut all these rows or columns.

6： if $\rho_k > h*\rho_0$
$S=H_k$ is the result hotspot, else $k=k+1$, goto **5**.

7： End

4. Hotspot Detection

The purpose of hotspot detection algorithm is to find the regions with relatively higher density of vessels activity. It have 6 basic step as shown in Algorithm 1.

Figure 2 shows an example of the cutting process in step 3. In this example, the whole number of vessels is 50, ρ_0=2, the whole region is splitted into 5*5 buckets,the density threshold ***h*** is 3,the cutting process stop after three iterations, and the result is two sub region with the density higher than 6.

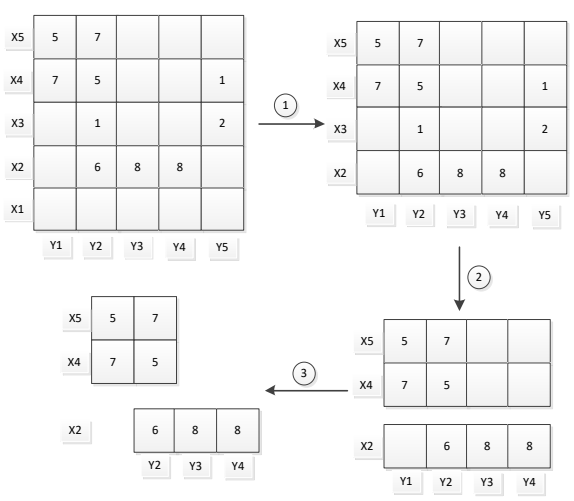

Fig. 2: The cutting process of hotspot area

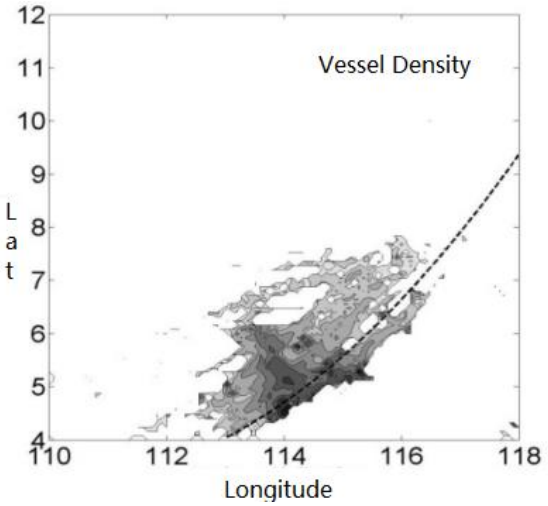

Fig. 3: The vessel density on experiment sea area

5. Experiments

5.1. Experiment preparation

The hadoop platform used consist of four pc nodes, one of them is the master node, the other three is the slave nodes, each node have one Intel Q8300 CPU with 4GB memory, the operation system on each node is Ubuntu Linux 14.04, the hadoop version is Hadoop1.2.1.

The data set used is a subset in the real AIS data on 2012 jan,the size of the whole data set is 102GB. After data preprocessing we choose a subset of vessel data in longitude 110ºE~118ºE， latitude 4ºN~12ºN, a part of south China Sea near Malaysia, and the dirty data and false data is already removed.

In this experiment, the whole region is splitted into 100*100 buckets,after the calculation in SpatialHadoop, the density of vessels activity is shown in figure 3.

5.2. Hotspots detection results

The figure 4 shows the relationship between threshold h with the hotspot area,when h increasing the hotspot area decrease greatly, but after the certain critical point shown in the figure, the hotspot area become stable.In this experiment, we split the data set into two sub set, one is the data on coastal area, the other is the high open sea area. The critical point of threshold h for two sub set are 100 and 39, which should be used in practice.

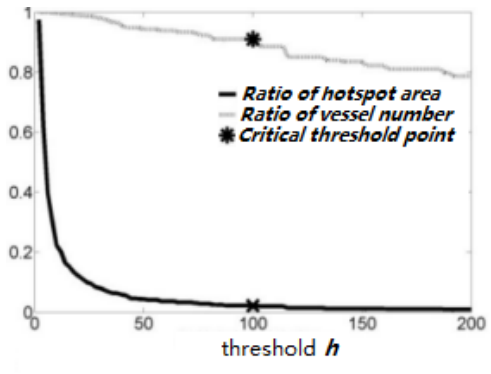

Fig 4a: coastal data set

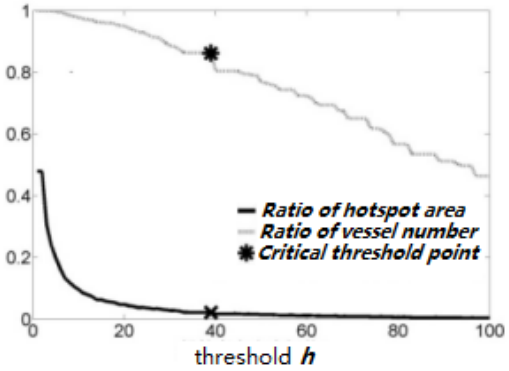

Fig4b: High open sea data set

We set the thresholds at ***h***=100and ***h***=39 for two data set, and the algorithm got the hotspot area show in figure 5.

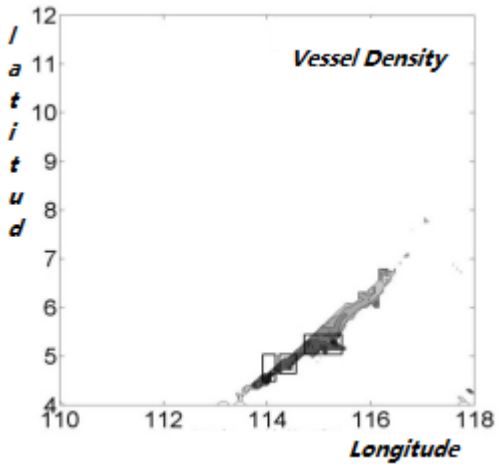

Fig 5a: Hotspot on coastal data set

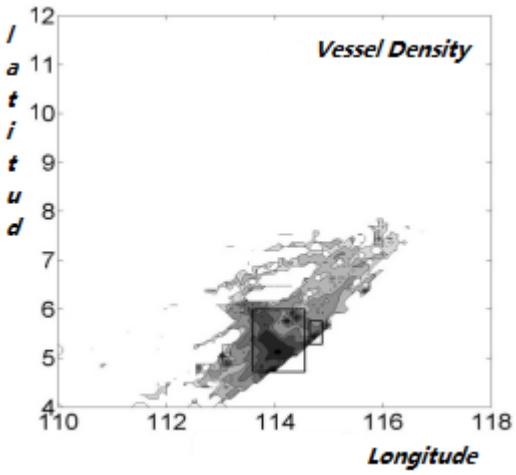

Fig5b: Hotspot on High open sea data set

5.3. Performance analysis

We test the performance of our algorithm with two naive hotspot computation approaches,one of which neither use the spatial index or the hadoop framework and working on the master node ,the other one only apply the hadoop framework but do not have spatial index. The results are shown in table 1.

Table 1: The Algorithm performance comparison

Time(seconds)	Grid 1*1	Grid 10*10	Grid 100*100
No R-tree index No Hadoop	735	75631	7692217

No R-tree index On Hadoop	12212	22121	78121
R-tree index On Hadoop	7372.88	7404.94	10001.15

6. Conclusion

The distribution of moving vessels' traffic data are usually heavily skewed large data sets,the spatial area and the size of data set are both very huge. The hotspot detection in this data is very helpful for decision making or further spatiotemporal data mining applications.In this paper, we proposed a grid-based hotspot detection algorithm on maritime traffic data. It split the spatial area into certain number of small buckets, and calculate the density on each bucket instead of compute the exact location of each single moving vessel. We apply the R-tree spatial index on Spatial Hadoop framework in the computation on the density of each bucket. Experiments on real AIS data shows that the method propose is effective and fast.

This work is just a first step,many challenges lies ahead. The algorithm proposed in this paper only takes the spatial attributes into computation. It can not use the temporal attribute and behavioral attribute well, these information can reflect the objects' customs, habits and even the characteristic of a certain region. Future research can be made on integrating the temporal and trajectory semantics of vessels with hotspot detection.

7. References

[1] Scrofani J W, Tummala M, Miller D, et al. Behavioral detection in the maritime domain[C]. *System of Systems Engineering Conference. IEEE*, 2015:380-385.

[2] Iphar C, Napoli A, Ray C. Detection of false AIS messages for the improvement of maritime situational awareness[C]. *Oceans. IEEE*, 2016.

[3] Liu B, Souza E N D, Matwin S, et al. Knowledge-based clustering of ship trajectories using density-based approach[C]. *IEEE International Conference on Big Data. IEEE*, 2015:603-608.

[4] Aji A, Wang F, Vo H, et al. Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce.[J]. *Proceedings Vldb Endowmen*t, 2013, 6(11):1009-1020.

[5] Yu J, Wu J, Sarwat M. GeoSpark: a cluster computing framework for processing large-scale spatial data[C]. *Sigspatial International Conference on Advances in Geographic Information Systems. ACM*, 2015:70.

[6] Eldawy A, Mokbel M F. SpatialHadoop: A MapReduce framework for spatial data[C]. *International Conference on Data Engineering. IEEE*, 2016:1352-1363.

[7] Juan L U, Tang G, Zhang H, et al. A Review of Research Methods for Spatiotemporal Distribution of the Crime Hot Spots[J]. *Progress in Geography*, 2012, 31(4):29-35