ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 689-696
doi: 10.18178/wcse.2019.06.102

A Hadoop-based Co-occurrence Pattern Mining Model on AIS data

Bao Lei [+]

Computer Science Department, Wuhan Donghu University, Wuhan,China

Abstract. AIS is a tracking and self-reporting system used by maritime vessels to exchange information with other ships, AIS base stations, and satellites.Co-occurrence mining in AIS data can measure the proximity of ships in space and time and can be used in maritime traffic monitoring or other security purpose.Most of the current existing approaches can not meet the practical needs of large-scale ship trajectory data mining due to the lack of designing on parallel computing architecture and are insensitive to the spatial data characteristics.A model on co-occurrence pattern mining based on Hadoop is presented in this paper.By using parallel partitioning on the original data set, the mining in ship trajectory data is implemented on an extended MapReduce architecture.The experiments on real AIS data sets show that the large-scale ship trajectory data can be processed effectively , and the efficiency and correctness are maintained.

Keywords: Spatiotemporal Data Mining; Spatiotemporal Co-occurrence; Apriori; AIS

1. Introduction

With the continuous extensive use on geographic information, the acquisition and calculation ability of ship trajectory data is constantly developing. Among them, the Auto Identification System is one of the most important maritime trajectory data source which can collect massive data each day. The AIS data set is vast and abundant, and the mining on it is a vital area of relative research. The motion pattern mining introduce trajectory pattern mining algorithm,which can find out valuable patterns in large trajectory data, so as to perceive the situation and obtain more economical and safe navigation information.Consequently, mining on AIS data has become an increasingly important research theme, attracting the attention from numerous areas, including computer science and geography.

While most of the current existing approaches can not meet the practical needs of large-scale ship trajectory data mining due to the lack of designing on parallel computing architecture and are insensitive to the spatial data characteristics.A model on co-occurrence pattern mining based on Hadoop is presented in this paper.By using parallel partitioning on the original data set, the mining in ship trajectory data is implemented on an extended MapReduce architecture.

2. Related Works

Spatiotemporal co-occurrence pattern mining [1] is a typical issue in motion pattern mining. Spatio-temporal co-occurrence pattern refers to two (or more) object instances that are adjacent in space and time.The co-occurrence pattern mining on ship trajectory data has great application value.For example, smuggling vessels may be in a co-occurrence mode for a long time during navigation.Ships often sail in formation during their maritime missions,different forms of formation organization mean different operational tasks.Pirates often follow up for long periods of time and at close range before unlawful actions are carried out. These situation can all be modeled as spatiotemporal co-occurrence patterns, and the corresponding mining method can be used for real-time discovery, which provides an effective means of analysis and calculation for military

[+] Corresponding author. Tel.: +027 81631688; fax: +027 81631687.
E-mail address: blnj2000@163.com.

operational plan and tactical action discovery, road and network plan in traffic field, and landmark event correlation analysis[2] in territorial defense.

At present, most of the related researches extend the spatial co-occurrence model to the time dimension. And the apriori algorithm is the main method used. Celik defines a measurement method of hybrid spatiotemporal co-occurrence pattern, proposes a hybrid spatiotemporal co-occurrence pattern algorithm, and makes an analysis of its correctness and completeness [3].The method is used to discover fixed patterns in animal migration [4] [5].[6] analyzed the mining on frequent co-occurrence sub-sequences between different moving objects from spatiotemporal data and proposed a two-stage mining algorithm, in that algorithm, the original trajectory is transformed into a sub-sequence with similar characteristics by hash function, and then the frequent fragments are mined by apriori algorithm.[7] combines the measurement of time dimension and space dimension and introduce a spatiotemporal co-occurrence pattern discovery algorithm COSTCOP. In this model the problems of time dimension and space dimension are considered comprehensively.

However, the huge amount of maritime location data has put forward new requirements for spatiotemporal co-occurrence pattern mining.At present, the amount of maritime location data has already reached the TB level, and it is increasing continuously.Taking AIS data as an example, the global AIS data collected by satellites and base stations is about 4 gigabytes a day, and the amount of AIS data generated in one year can reach 1.4 TB.Most of the existing research results are based on a small data set, which can not cope with the huge amount of maritime data. Moreover,the speed of co-occurrence pattern mining is limited by the serial computing mode of traditional mining algorithms and the insensitivity of spatial features.To solve these problems, this paper proposes a spatio-temporal co-occurrence pattern mining algorithm based on spatial Hadoop analysis platform architecture for large maritime location data.By dividing the original data set with two-level spatial index, the spatiotemporal co-occurrence pattern mining in ship trajectory data is realized on the extended MapReduce architecture.The experimental results based on real AIS data sets show that the large-scale ship trajectory data can be processed effectively, and the efficiency and correctness are maintained.

3. Spatiotemporal Co-occurrence Pattern Model

Ship spatiotemporal co-occurrence pattern refers to the ship location satisfies certain neighborhood relationship in position and time marking.This neighborhood is determined by the degree of spatiotemporal proximity participation of an instance.

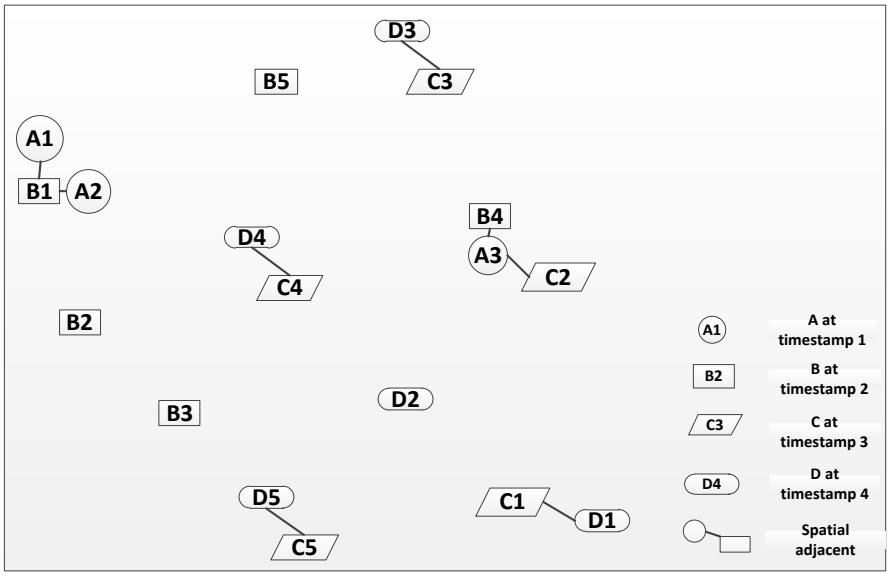

Fig. 1: The Sample graph of ship Spatio-temporal co-occurrence pattern

3.1. Spatiotemporal adjacent

Spatiotemporal adjacent includes spatial adjacent temporal adjacent.Spatial proximity is measured by Euclidean distance.The time proximity relationship is measured by the difference of the time stamp of the ship

position data instance. In Figure 1, the spatial adjacent relationship is the instances connected by solid lines.Instances that satisfies both spatial and temporal adjacent are A1B1,C1D1,C3D3,C4D4,C5D5.

3.2. Support

Support refers to the ratio of adjacent instances in a data set.When the ratio of adjacent instances is more than a certain threshold, it is determined as spatiotemporal co-occurrence pattern.

3.3. Confidence

CPR is the confidence of spatiotemporal co-occurrence pattern, and its formula is

$$CPR_i = \frac{\text{Number of } f_i \text{ instance in } C_k}{\text{Whole Number of } f_i \text{ Instance}} \quad (1)$$

where Ck={f0,... Fk-1} is the k-element candidate spatiotemporal co-occurrence pattern, fi ∈ E, E is the complete set of spatiotemporal objects. Table 1 shows the CPR in figure1.

Table 1: The instance participation rate

Candidate set	AB		CD	
	A	B	C	D
CPR	1/3	1/5	3/5	3/5

Various combinations among ship objects constitute candidate spatiotemporal co-occurrence patterns. If it still satisfies the confidence condition, it will be a space-time co-occurrence model.

$$CPR(C_k) > \theta \quad (2)$$

Among them, θ is the confidence threshold. the candidate space-time co-occurrence mode {C, D} in Table 1 is a binary space-time co-occurrence pattern when the threshold is 0.5.

4. Co-occurrence Pattern Mining Algorithm Design on Hadoop Architecture

The vital problem of spatial-temporal co-occurrence pattern mining algorithm based on Hadoop is data partitioning and parallelization of mining algorithm.

4.1. Data partitioning

The distribution of ship position data is heavy skewed and in-homogeneous.The uniform data partitioning result in unbalanced data load, which will lead to the decrease of parallel computing speed and affect the overall efficiency of the algorithm.At the same time, data partitioning must take spatial locality into account, that is, objects close in space should be partitioned into the same partition.

The algorithm have three steps:

Step1: Calculating the partition number N

Partition Number $N = \lceil S(1+\alpha)/B \rceil$,among them, S is the size of the input data file and B is the capacity of HDFS block, representing the overhead rate. It is mainly used to save duplicate spatial records and store local indexes.

Step2: Determining partition boundaries

In partition, the input files are randomly sampled in order to simplify the calculation.The default sampling rate for input files is 1%.When the input file is large, the MapReduce task is constructed to browse all records distributed and output 1% of the sample files.For the sample file, the minimum outer rectangle (MBR) of each data block is saved to form a set of boundary partition rectangles.

Step3: Partitioning the data set

It is a matching problem between original data and partitioned area to implement data partition operation according to the known MBR boundary.Because the matching of partitions between the original data is independent with each other, it can be processed in parallel.Firstly, the original data is divided into n blocks on average, and a Map process is established for each block of data, then the execution results are fed into Reduce process for simplification, thus speeding up the processing.

4.2. Co-occurrence pattern ming algorithm

Spatiotemporal co-occurrence pattern mining algorithm is essentially an adaptive improvement of classical Apriori algorithm[8-10] for large data set of ship location data. The algorithm include data partitioning, parallel mining of partitioned data, eliminating boundary common reality examples in the results, and then executing Apriori mining algorithm, as shown in figure 2.

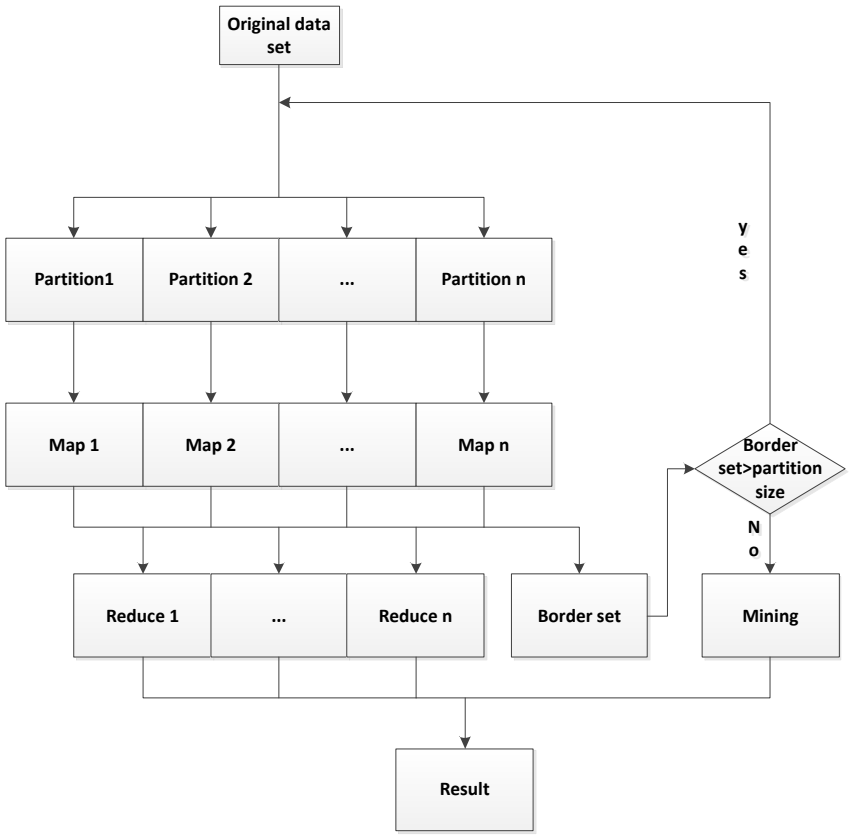

Fig. 2: Iterative parallel mining process

Parallel computation of boundary determination algorithm only involves parallel sorting and horizontal cutting of vertical slices.For the final partion result, the rectangular area range of each partition is required.

The partition function use <rectangle，records> as parameters，rectangle is the Minimum Boundary Rectangle of current data set,records is the data set itself.

ALGORITHM 1 THE SPLITTER ALGORITHM

function Splitter

Input:<rectangle,records>
Output:<null,rectangle i>

```
  p=records
While(p.next!=null)
  do sequence by latitude
  update result
end
for i=1:1:
  Q=result(i to i+records/     )
  rectangle i=(Q_lon_min,Q_lat_min,Q_lat_max,Q_lat_max)
  output <null,rectangle i>
  end
```

The mining algorithm use Map_Reduce calculation in spatiotemporal co-occurrence patterns mining.The Map function is to compute spatiotemporal co-occurrence instances. It have 2 parameters, rectangle is the Minimum Boundary Rectangle of current data set, ArrayWritable is the data blocks.

ALGORITHM 2 THE MAP FUNCTION FOR MINING

Map function of Co-occurrence mining
Input：<rectangle，ArrayWritable>
Output：<AISWritable，IntWritable>
For each ArrayWritable p in rectangle For each ArrayWritable q in rectangle If p and q meet nearby relation R If (p and q) ∉ Border area Output<AISWritable(p,q)，one> End End End End

The Reduce function of co-occurrence patterns mining function is to merge the same spatiotemporal co-occurrence instances.

ALGORITHM 3 THE REDUCE FUNCTION FOR MINING

Reducefunction of Co-occurrence mining
Input：<AISWritable，IntWritable（1）>
Output：<AISWritable，IntWritable>
For each AISWritable m For each ArrayWritable n If m = n IntWritable= IntWritable+1 Delect ArrayWritable n end End End Output<AISWritable，IntWritable>

5. Experiments

5.1. Experiment evironment and data preparation

Four computers are used to build Hadoop platform. Each computer serves as a computing node, one of which serves as Master and JobTracker control nodes, and the other three serves as Slave and TaskTracker computing nodes.The operating system is Ubuntu Linux 14.04, and Hadoop version is Hadoop 1.2.1. The basic configuration of the computer is InterQ8300, 4G of memory and 500g of hard disk.

The experimental data were selected from real AIS historical data. The global AIS data of January and July 2012 are selected, with a data volume of about 400 GB. The selected sea area is near the Qiongzhou Strait in China.Before the experiment, the data were preprocessed to eliminate erroneous data.

5.2. Experimental results

(1)January 2012 Data Set

Experiment parameters: input data = 224 GB, block size = 1 MB, adjacent proximity = 500 m, support = 10 000, confidence = 0.1. The experimental results are shown in Table 2.

Table 2: The results of data set 2012 Jan.

MMSI-1	MMSI-2	Adjacent	Support	Confidence
413376050	413591670	40	56810	0.4034501
412044710	412764290	3	18605	0.28056974
412044710	412044770	50	25785	0.23463254
412044770	413354380	3	13197	0.14927635
412379030	413354380	5	12368	0.14149418
412044750	667003097	3	24649	0.110755
412044710	413040140	10	45592	0.10729858
412044710	413376050	16	55848	0.10295803

(2)July 2012 Data Set

Experiment parameters: input data = 186 GB, block size = 1 MB,adjacent proximity = 500 m, support = 10 000, confidence = 0.2. The experimental results are shown in Table 3.

Table 3: The results of data set 2012 July

MMSI-1	MMSI-2	Adjacent objects	Support	Confidence
224422360	224943140	2422	19628	0.716442308
413373960	413374010	541	12147	0.31855385
413376550	413376560	1344	43577	0.27512794
413800901	413816844	566	20126	0.23038098
412036350	413443970	509	15645	0.22598118
412011360	412011370	1917	58825	0.2143528
412001530	412207820	565	21056	0.20106762
412018160	412018180	1623	80465	0.20043223

5.3. Analysis

The co-occurrence pattern mining algorithm has four controllable parameters: partition size, adjacent proximity, support and confidence.We analyses and discusses the influence of each parameter on the efficiency of the algorithm.The data used is January 2012, about 224 GB.

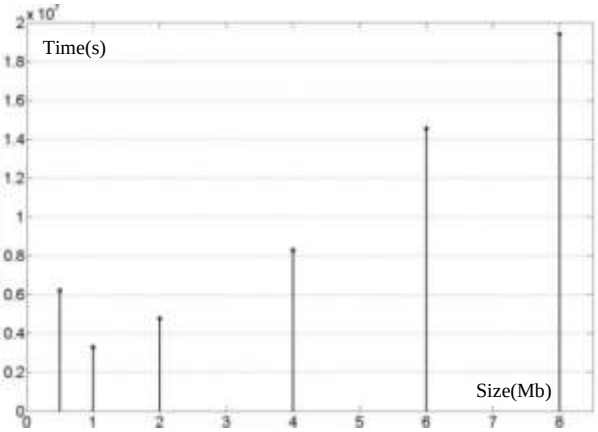

Fig. 3a: Partition size analysis.

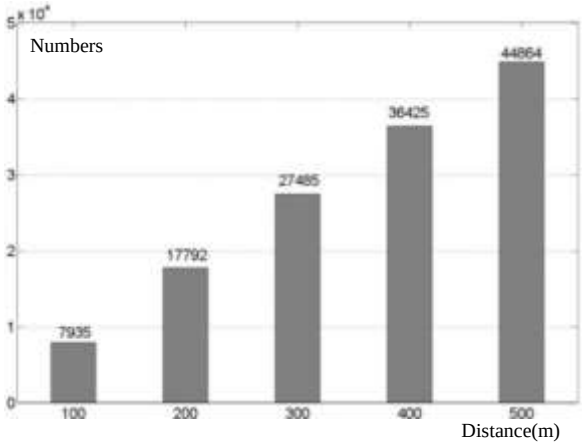

Fig. 3b: The adjacent distance threshold analysis.

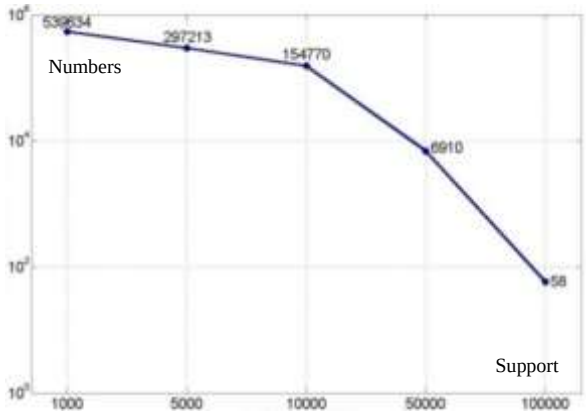

Fig. 3c: The Support threshold analysis,

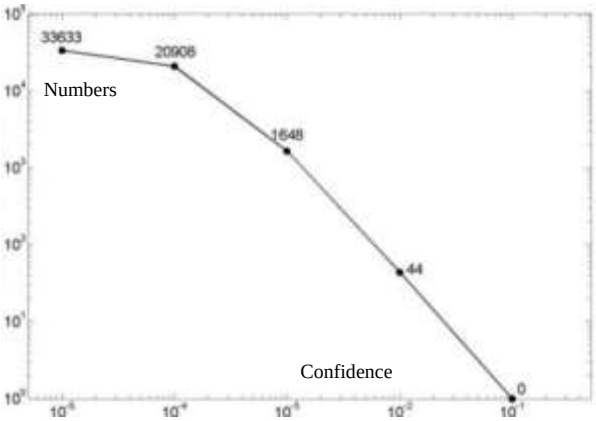

Fig. 3d: The confidence threshold analysis

As shown in Figure 3, with the increase of partition size, the execution time of the algorithm decreases first and then increases, and the minimum value is obtained when the partition size is 1MB.The adjacent proximity is a very important parameter of spatiotemporal co-occurrence model, which directly affects the reliability of the experimental results. If the adjacent proximity threshold is too large, the spatial-temporal co-occurrence relationship will be weakened. If it is too small, some spatial-temporal co-occurrence patterns will be neglected. The increase of support by one order of magnitude will decrease the number of ajacent objects by four orders of magnitude. And the degree of confidence reflects the reliability of space-time co-occurrence relationship. The spatial-temporal co-occurrence model with strong spatial-temporal co-occurrence relationship is of practical significance. With the increase of confidence parameters, the number of common cases decreases rapidly

6. Conclusion

Spatio-temporal co-occurrence pattern mining we present here is on ship AIS data, by calculating the spatio-temporal relationship between objects in massive AIS data, finds the patterns and rules behind the co-occurrence relationship, which has very important practical value for military and civil fields such as marine traffic and safety supervision.

The algorithm proposed is a Hadoop-based one. By using parallel partitioning algorithm to divide the original data set, we implement spatio-temporal co-occurrence pattern mining in the extended MapReduce framework, and carry out experiments and analysis based on the actual data set of AIS.

Due to the limitation of conditions, the scale of the experimental data selected in this paper is still small. It is also an important research direction to use larger global data for experiments and combine other data sources to comprehensively to judge ship abnormal behavior.

7. References

[1] Liu DY,Chen HL,Qi H, et al. A survey on spatiotemporal data mining[J], Research and Development on

Computers.2013,50(2):225-239.

[2] Mete C,Shakhar S,James P,et al.Mixed- Drove Spatialtemporal Co-Occurrence Pattern Mining[J].IEEE Transactions on knowledge & data engineering,2008,20(10):1322-1335.

[3] Mete C. Discovering Partial Spatial-Temporal Co-occurrence Patterns[C]. IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services,Icsdm 2011,Fuzhou,China,June 29-July.2011:116-120.

[4] Mazzarella F,Vespe M,Damalas D,et al.Discovering Vessel Actives at Sea using AIS Data:Mapping of Fishing Footprints[C].International Conference on Information Fusion.2014:1-7.

[5] Clement I,Aldo N,Cyril R.Detection of false AIS messages for improvement of maritime situational awareness[C].Oceans'2015,Oct 2015,Washington,DC,United States.

[6] James W,Muraki T,Donna M,et al.Behavioral Detection in the Maritime Domain[C].The 10th Systems Engineering Conference(SoSE), 2015,380-385.

[7] Liu B,Matwin S.Knowledge-based Clustering of Ship Trajectories Using Density-based Approach[C].IEEE International Conference on Big Data,2014.

[8] He FZH. Research on Spatial Co-location Pattern Mining Based on Parallel Computing[D]. University of YuNan,2014.

[9] Hao XF, Tan YSH,Wang JY. Parallelization Research and Implementation of Apriori Algorithms on Hadoop Platform [J].Computer and Modernizatio,2013,(3):1-4.

[10] Zhu AZH.Research on Enhancing Apriori Algorithms Based on Hadoop.Huazhong University of Science and Technology,2012.