PROCEEDINGS OF SCIENCE

# Framework of Frequently Trajectory Extraction from AIS Data

**Xiaoyu Jin**[1]
*State Key Laboratory of Networking and Switching Technology,Beijing University of Posts and Telecommunications*
*Beijing,100876, China*
*E-mail:* `jinxiaoyu@bupt.edu.cn`

**Yang Yang**
*State Key Laboratory of Networking and Switching Technology,Beijing University of Posts and Telecommunications*
*Beijing,100876, China*
*E-mail:* `yyang@bupt.edu.cn`

**Xuesong Qiu**
*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications*
*Beijing,100876,China*
*E-mail:* `xsqiu@bupt.edu.cn`

Mining trajectory data has been attracting significant interest in the last years. Emerging technologies like Automatic Identification System (AIS) provides multi-dimensional data which is about voyages and vessels. The maritime area is a free moving space. Unlike the vehicles' movements are constrained by road networks, there is no such a sea route for ships to follow in maritime area. In this paper, we propose a framework of frequently voyage extraction from AIS data, which learns frequently voyages using improved dynamic time warping(IDTW) and Adaptive Density Increment Clustering (ADIC). ADIC can adaptively adjust parameters on uneven density data . We conduct the experiments on real maritime trajectories to show the effectiveness of proposed framework.

*CENet2017*
*23-24 July, 2017*
*Shanghai, China*

[1] The work presented in this paper was supported by NSFC(61401033), National Science and Technology Pillar Program Project (2015BAI11B01).

## 1.Introduction

Automatic Identification System (AIS) technology provides a vast amount of near-real time information. Trajectory data mining is a challenge task attracting significant interest. The trajectories of the vessels is more complicated because the vessel's sailing space is freer than vehicle’s. A ship’s movement may not exactly repeat the same trajectory even the ship has the similar movement behavior with others. So the frequent trajectory extraction of the vessels is more difficult. This paper proposed a framework of frequent voyage extraction from AIS data using improved dynamic time warping (IDTW) and Adaptive Density Increment Clustering (ADIC). In the future the result of frequent voyage information can used in vessel trajectory prediction and trajectory anomaly detection.

The novel contributions in this paper are as follows:(1) We use improved DTW to measure the similarity of two trajectories. The improved DTW(IDTW) can limit the slope of the matching route.(2) The algorithm of Adaptive Density Increment Clustering(ADIC) is proposed to adapt the trajectory data with uneven density.

The rest paper is organized as follows. Section 2 introduces some related works. The section 3 proposes our framework to mining frequent trajectories from ais data. We compare theperformances of algorithms proposed in this paper and the existing algorithms in section 4. Section 5 concludes the paper.

## 2.Related Works

In data mining field, unsupervised methods have been extensively researched recently. As an important unsupervised method, clustering has been paid more attention. Collections of clustering analysis algorithms are proposed in the following paper. K-means, DBSCAN, OPTICS, and STING are described respectively[1-4].The key idea of DBSCAN[2] is that for each object of a cluster the neighborhood of a given radius(*Eps*) has to contain at least a minimum number of objects(*Minpts*). There are two given parameters. However, the density of clusters is usually different, parameters have great influence on clustering results. We propose ADIC in this paper to solve the problem of clusters with different density.

The spatio-temporal trajectory is spatial position data set based on time series.There are also many method of similarity measure for trajectory. DTW [5] is usually used to calculate the similarity between two sequences, which is a kind of dynamic programming method for time series similarity measure. LCSS distance can measure the similarity of the trajectories by obtaining the longest common subsequence between the trajectories[6].

An unsupervised and incremental learning approach to the extraction of maritime movement patterns is presented to convert from raw data to information supporting decisions[7]. A methodology that aims to convert the large amount of AIS data into decision support elements, independently of the number of receivers, their performance, the platform of origin and the scale of the area of interest.Giuliana et al extracted motion patterns which are then used to construct the corresponding motion anomaly detectors from the real historic AIS data[8]. Adaptive kernel density estimation is used to detect abnormal trajectories, which needs large computation.

## 3.Framework of Frequent Voyage Extraction

The framework of frequent voyage extraction proposed in this paper includes three steps, data sampling, similarity measure of trajectories and clustering meathod. Method of data sampling is described as 3.1. An improved DTW method is used to calculate the similarity between trajectories. Adaptive density increment clustering as the clustering meathod gathers similar trajectories.

### 3.1data sampling

AIS data includes static information(e.g.,vessel type,destination, cargo type, length, width),and dynamic information(e.g.,time stamp,longitude,latitude,course over ground, speed over ground). The time interval between two vessel location point in the AIS data is inconsistent, changing from a few seconds to a few minutes, so the AIS data should be resampled in equal time interval. We set the time interval equal to 1min.

If the point *P* at time *t* is needed,we find the nearest point $P_1$ earlier than *t*,and the nearest point $P_2$ later than *t*. the point $P_1$ is at time $t_1$, $P_2$ is at $t_2$.The attributes of point includes latitude, longitude, course and speed. The attribute value of the point *P* is as formula(1)：

$$value_p=\frac{(value_{p2}-value_{p1})}{t_1-t_2}*(t-t_1)+value_{p1} \tag{3.1}$$

*Value* can represent any value of attribute in trajectory point.

### 3.2Describe of improved Dynamic Time Warping

AIS can be regarded as spatio-temporal sequence data.Dynamic time warping (DTW) can capture the dynamics and match patterns for different time series. In the DTW algorithm, there is no limit in the matching path of the trajectory points, which may lead to a great difference in time between the matching trajectory points.

Assume $T_1$ and $T_2$ are two trajectories to be compared. The first step of IDTW is to define the distance between points of the two trajectories. We use euclidean distance in this paper. Location information in AIS data is longitude and latitude. Formula(3.2) and (3.3) is used to calculate the distance between two points.

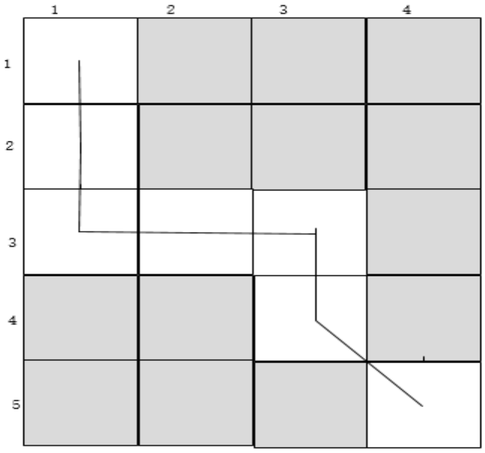
$$C=\sin^2\left(\frac{lat_i-lat_j}{2}\right)+\cos(lat_i)*\cos(lat_j)*\sin^2\left(\frac{lon_i-lon_j}{2}\right) \tag{3.2}$$

$$D(T_1(i),T_2(j))=\arcsin(\sqrt{C})*2*R \tag{3.3}$$

$T_1(i)$ is a point of trajectory $T_1$, $T_2(j)$ is a point of trajectory $T_2$. $D(T_1(i),T_2(j))$ is the distance between $T_1(i)$ and $T_2(j)$.*R* is the Radius of the earth.The algorithm of DTW can be described as fomula(3)：

$$DTW(i,j)=D(i,j)+min(DTW(i-1,j),DTW(i,j-1),DTW(i-1,j-1)) \tag{3.4}$$

The aim of DTW method is to search the minimum total length of the accumulate path which can be achieved by formula (3.3)[5]. If the point $T_1(i)$,$T_2(j)$ is in the optimal path, then the path from point $T_1(1)$, $T_2(1)$ to $T_1(i)$, $T_2(j)$ is also locally optimal. Result of similarity measure is $DTW(T_1(i), T_2(j))$, $T_1(i)$ and $T_2(j)$ are the last point in *T1* and *T2* respectively. The matching path is as Figure1,(*$T_1$(1)T2(1)*, *$T_1$(1)T2(2), $T_1$(2)T2(2), $T_1$(3)T2(2), $T_1$(3)T2(3), $T_1$(3)T2(4), $T_1$(4) T2(5))*. However, there is no limit in the matching path of the trajectory points, which may lead to a great difference in time between the matching trajectory points.



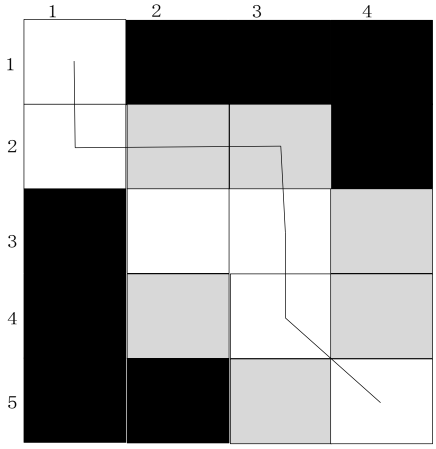


**Figure 1:** matching path in DTW **Figure 2:** matching path in IDTW

In order to conquer the shortcoming of the DTW algorithm mentioned above, this paper proposes the improved DTW(IDTW),which uses the speed attribute in AIS data to constrain the matching path. The IDTW is as follows:

(1) Before calculate the distance of between *$T_1(i)$* and *$T_2(j)$*, we should calculate the average speed between the start of *$T_1$* and point *$T_1(i)$*. Similarly the average speed for *$T_2(j)$* also should be computed.

Define the average speed of *$T_1(i)$* as following:

$$SP(i)=\frac{\sum_{x=1}^{i} SOG_{px}}{i} \tag{3.5}$$

$$D(T_1(i),T_2(j))=max \quad \text{if} \quad \frac{j}{i}<\frac{w_2 * SP(i)}{SP(j)} \quad \text{or} \quad \frac{j}{i}>\frac{w_1 * SP(i)}{SP(j)} \tag{3.6}$$

If *i* , *j* meets the condition in formula(6), the distance between *$T_1(i)$* and *$T_2(j)$* is as formula(6). If not, the distance is calculated using formula(3.2) and formula(3.3). In this paper, we set $w_1$=2,$w_2$=0.5, *max* is equal to the largest distance between *$T_1(i)$* and *$T_2(j)$*.

(2) Calculate the similarity between *$T_1$* and *$T_2$* using formula(3.4).

The constraint matching path is shown in Figure 2,the black squares represents the value of *max*, the path does not choose these black squares.

### 3.3Adaptive Density Increment Clustering (ADIC)

There is uneven distribution of trajectory's density in the detection region. The traditional DBSCAN treats the clusters with small density as noise, which makes partial information of frequent trajectories loss. The trajectory data in the detection area is updated constantly. When a new trajectory arrives, the cluster result is completely recalculated, which leads to a large number of computation. ADIC proposed in this paper uses adaptive parameters to deal with trajectory clusters with different densities. When a new trajectory arrives, only affected trajectories updates their cluster results in ADIC.

### 3.3.1Definition of Parameters

Calculate the similarity between trajectories using IDTW. If $DTW(T_i,T_j)< r$, $T_j$ is called $T_i$ 's neighbor. $\epsilon$ is The minimum number of neighbor points. The number of $T_i$'s neighbor is $N(T_i)$.

if $N(T_i)< \epsilon$ , $T_i$ is considered as noisy point. If not, compute the $d(T_i)$ distance between the $\epsilon$-th nearest neighbor trajectory and the $T_i$ ,as shown in Figure3.$\epsilon$=3,$N(T_i)$=5.

Then use formula (3.7) to compute the $dr(T_i)$,$T_j$ is $T_i$'s neighbor,$dr(T_i)$ is the ratio of $d(T_i)$ and the average value of $d$ for $T_i$'s neibghbor. $n$ is the number of non-noise trajectories in $T_i$ 's neibghbor. If $n$=0,$T_i$ is considered as noisy point too. If $dr(T_i)>dr\_value$, $T_i$ is considered as core point, otherwise $T_i$ is a boundary point.

$$dr(T_i)=\frac{d(T_i)*n}{\sum_{j=1}^{n} d(T_j)}(n\geqslant 1) \tag{3.7}$$

### 3.3.2Clustering Rules

Set the adaptive radius of core point $T_i$ as formula(3.8):

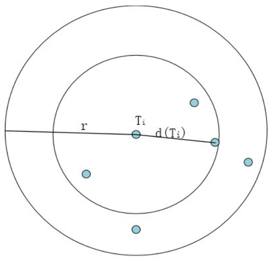
$$R\,var=\sqrt{r*d(T_i)} \tag{3.8}$$

Value of $Rvar$ is depended on $r$ and $d(T_i)$, if point density around $T_i$ is large, then $d(T_i)$ must be small, and $Rvar$ commensurately reduces.

There are three definitions as follow:

- Direct density reachable： if $d(T_i,T_j)<Rvar(T_i)$ and $T_i$ is a core point, $T_j$ is not a noisy point, $T_i$ and $T_j$ are direct density reachable.
- Density reachable：If there are a series of trajectories $T_i$ $T_1$, $T_2$, $T_3$…$T_j$, and $T_i$, $T_1$,$T_2$,$T_3$… are core points, $T_j$ is not a noisy point, the adjacent points are direct density reachable,$T_i$ and $T_j$ are density reachable.
- Density connected： If there are two series of trajectories $T_i$, $T_{11}$, $T_{21}$, $T_{31}$...$T_{j1}$, and $T_i$, $T_{21}$,$T_{22}$,$T_{23}$...$T_{j2}$, $T_i$, $T_{11}$, $T_{21}$, $T_{31}$…and $T_{21}$,$T_{22}$,$T_{23}$… are core points, $T_{j1}$ and $T_{j2}$ are not noisy points, adjacent points are direct density reachable, $T_{j1}$ and $T_{j2}$ are density connected.

The clustering rule is: density connected points belong to the same cluster, as the Figture4 shown.



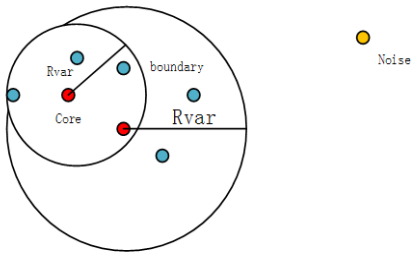


**Figure 3:** trajectory Tᵢ and its neighbor **Figure 4:** clustering rules

## 4.Experiment Analysis

To analyze the performances of the framework proposed in this paper, we conduct similarity measure of trajectories based HMM [9] and the trajectory clustering algorithm imoroved DBSCAN [7]. HMM determines the similarity of trajectories using the conformity of the corresponding HMM models. A HMM model is based on a trajectory. The probability that a HMM model trained on one trajectory generates data of the other trajectory is the similarity between the two trajectories. The weakness of the HMM similarity algorithm is that it takes long time to generate and select the best model for each trajectory. Experimental environment is as follows: memory 8G, CPU 2.39GHz, python 2.7.In this experiment the computation time of two algorithms is compared. We conduct the experiment on trajectories with different lengths. The result is as Figure5.

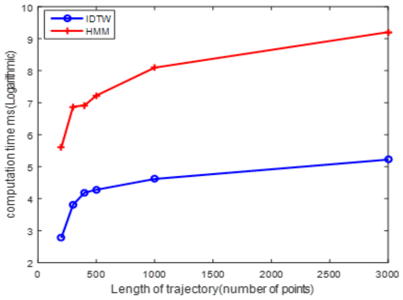


**Figure 5:**computation time of IDTW and HMM

We use millisecond to measure computation time in Figure5.The logarithmic scale of the figure allows the full range of computation time to be displayed. As shown in Figure5, the computation time of HMM similarity algorithm increases rapidly as the trajectory length increases. When number of points increases by 10 times, the computation time of IDTW increases by 5 times, and HMM increases by 10 times. So the IDTW proposed in this paper is more suitable for similarity measure of long trajectories. When using IDTW on a trajectory with 500 points, the computation time is reduced by 90% compared with HMM.

In order to demonstrate the performance of ADIC and imoroved DBSCAN[7], data set iris (http://archive.ics.uci.edu/ml/) is used in the experiment. The change of parameters has little effect on the clustering results of ADIC. In addition the density of clusters also has little influence on the clustering results in ADIC. The Clustering accuracy of improved DBSCAN and ADIC with different parameters is as Figure 6.

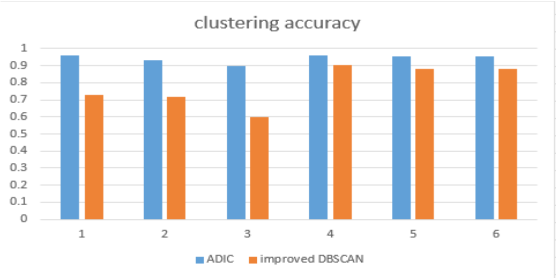


**Figure 6:** clustering acccuracy of ADIC and improved DBSCAN with different parameters

In Figure 6, the horizontal ordinate represents a series of different parameters. Parameters are as follows:(r=0.2,ε=4),(r=0.2,ε=5),(r=0.2,ε=6),(r=0.3,ε=7),(r=0.3,ε=8),(r=0.3,ε=9).

The maximum difference of the cluster accuracy is 0.06 in ADIC, so the change of parameters has little influence on the clustering accuracy of ADIC. However, the maximum difference in improved DBSCAN is 0.306. The clustering accuracy of improved DBSCAN is closely related to the clustering parameters.

We conduct the framework proposed in this paper on vessel trajectories.Our experimental data use AIS information collected from the west coast of North America. We select voyages for Seattle as clustering data. First, interpolation using formula (1) applies on trajectories. Then the IDTW algorithm is used for similarity measurement on trajectories. since the starting location for each voyage is not the same, we match these trajectory points from the last point of each voyage. Last, the ADIC algorithm is conducted on the trajectories.

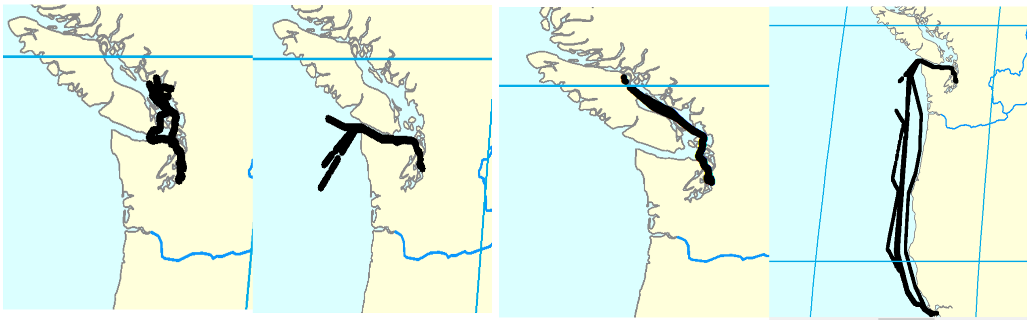


**Figure 7:**four cluster of trajectories around Seattle

As shown in Figure7, there are four types of voyages around Seattle. Frequent trajectories of voyages are discovered and classified using the framework proposed in this paper.

## 5.conclusion

The paper proposed a framework for frequent trajectories discovery. First, vessel trajectories should be interpolated and resampled. Then, two improved algorithm proposed in this paper are applied on these trajectories. The performance of IDTW and AIDC is showed in experiment results.

## References

[1] S Lloyd . *Least squares quantization in PCM*[J]. IEEE Transactions on Information Theory, 1982, 28(2):129-137.

[2] M Ester, HP Kriegel, J Sander, et al. *A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise*[C]// Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996:226-231.

[3] M Ankerst, M M Breunig, HP Kriegel, et al. *OPTICS: ordering points to identify the clustering structure*[J]. Acm Sigmod Record, 1999, 28(2):49-60.

[4] J B Bocca,M Jarke, C Zaniolo. *Proceedings of the 23rd International Conference on Very Large Data Bases*[J]. Morgan Kaufmann Publishers Inc, 1997, volume 49(11):10-11.

[5] C Hu, Q Zhao, N Luo. *Generalized trajectory fuzzy clustering based on the multi-objective mixed function*[J]. Journal of Intelligent & Fuzzy Systems, 2015, 29(6):2653-2660.

[6] X Gong, T Pei, J Sun, et al. *Review of the Research Progresses in Trajectory Clustering Methods*[J]. Progress in Geography, 2011, 30(5):522-534.

[7] G Pallotta, M Vespe, K Bryan. *Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction*[J]. Entropy, 2013, 15(6):2218-2245.

[8] B Ristic, B L Scala, M Morelande, et al. *Statistical analysis of motion patterns in AIS Data: Anomaly detection and motion prediction*[C]// International Conference on Information Fusion. IEEE, 2008:1-7.

[9] F Porikli. *Trajectory distance metric using hidden markov model based representation*[C]// ECCV PETS Workshop. 2004.