PROCEEDINGS OF SCIENCE

Mining Spatio-temporal AIS Data Using Prefix-span

Fumin Sun [1]
Institute of Software, Chinese Academy of Sciences, 4# South Fourth Street, Zhongguancun, Beijing 100190, PR.China
E-mail: sunalive01@163.com

Yong Deng
Institute of Software, Chinese Academy of Sciences, 4# South Fourth Street, Zhongguancun, Beijing 100190, PR.China
E-mail: 171166789@qq.com

Feng Deng
Institute of Software, Chinese Academy of Sciences, 4# South Fourth Street, Zhongguancun, Beijing 100190, PR.China
E-mail: dengfeng19901003@126.com

The prefix-span algorithm utilized for mining sequence data can guarantee the timescale for the data statues and has a great improvement in execution efficiency when compared with FP-GROWTH, which has a similar function. The spatio-temporal data is one kind of data that contains location and time stamp information together. This paper discusses the feasibility by using Prefix-span algorithm mining spatio-temporal data. This paper conducts a simulation experiment that mines the sequence Spatio-temporal data by using Prefix-span algorithm. The paper imports AIS data, one of the spatio-temporal data containing vessel motion information, to conduct the experiment. The results show that prefix-span algorithm can discover the crucial statues for spatio-temporal data with statues strictly sequenced according to their timestamp. This method can strongly support further researches in terms of anomaly detection of spatio-temporal data and the intelligent management of moving objects such as maritime traffic etc.

ISCC 2015
18-19, December, 2015
Guangzhou, China

[1]Speaker

PoS(ISCC2015)005

1.Introduction

In the data mining field, the sequential pattern mining used to mine frequent patterns from spatio-temporal datasets has become the most intuitive and attractive[1]. As the trajectory data stores the objects' location information in a period of time, we can reconstruct the issue of mining vessel motion patterns as extracting frequent sequential patterns in trajectories [2]. For the multiple dimension and high volume of spatio-temporal data, the traditional sequential pattern methods have complex computations and the timestamp of the spatio-temporal data cannot be strictly guaranteed.

The prefix-span algorithm is proposed by Han Jiawei to solve the problems for mining sequential data. The algorithm avoids testing every possible combination of a potential candidate sequential pattern [3]. It fixes the order of the items previously. There are many examples in the applications of prefix-span. One paper discovers frequent patterns in the wind speed and direction[4]. Ying J J C predicts the location of the objects using prefix-span analyzing the motion patterns based on the GPS data[5]. R Assam proposes a method for traffic management control based on the prefix-span algorithm[6]. Besides, on the trajectory data mining, the applications of the prefix-span extend to the behavior in life. Xin-zhi W detected the software behavior trustworthiness based on the prefix-span, as an example in security field[7]. And R.U magandhi used prefix-span to discover the user's habits from URL log in the search engine[8]. In this way, the recommended content is extracted.

In this paper, the application of the prefix span focuses on the vessel motion, which is still on account of the spatio-temporal data. The spatio-temporal data, one type of data, records the location and other information of vessel. AIS data is imported to conduct the experiment. In one paper , we have already introduced what is AIS[10], which is one of the self-reporting systems that has been led into trace any vessel and maritime surveillance. It has been proposed by International Maritime Organization and International Convention for the Safety of Life at Sea (SOLAS)[10]. The AIS facility is compulsorily required to be installed in many types of civilian vessels. In recent years, AIS data is the main source of information for maritime surveillance. By means of analyzing the attributes of the AIS data such as latitude and longitude, we can get the vessel motion patterns, predict the motion status and make detection of anomaly motion status[10]. Feixiang Z used the association rules to discover the association positions of the vessel[9], and gave the example of detecting anomaly through his method. In the paper, the time scale of tracks was considered [10-11].

The usage of the prefix-span will guarantee the time scale of the vessel trajectory. The division of the raw data is extended to get more information about the motion. The combination of the AIS data and the prefix-span is an innovative application of the algorithm.

In the first section of this paper, a simple back-grounds introduction is made; in the second section, the algorithm of prefix-span is introduced and then we will give out important arguments. Furthermore, we will combine the algorithm with the AIS data in the third section. The dimension of the data is extended in comparison with method in Ref. 9, which also applied the association rules to the AIS data. In the fourth section, we will discuss the results and give examples for further use of the method. In the fifth section, we will make a conclusion.

2. The Prefix-span Algorithm

The algorithm was proposed by Han Jiawei in 2004 aiming at discovering the order patterns from sequences. There are several kinds of methods to do the sequence mining, such as GSP and Free span, but they spent much time in finding candidates or costed highly in handling projected databases. To avoid checking every possible items, the order of items within each element can be fixed at first. For example, the sequence listed as <a(abc)(ac)d(cf)> instead of <a(bac)(ca)d(fc)>. The order of the sequence is defined as unique, thus the combination of different candidates are reduced. In the generation of a projection database, the prefix span projects only the possible subsequence and the projected sequence shrink rapidly. Under the situation that the raw dataset is huge, the prefix span can still have enough efficiency.

For better understanding of prefix span, some concepts are introduced , as shown in Ref.3; but the definitions in this paper are more adaptive to the spatio–temporal data.

Definition 1 (Prefix)[3]**.** Assume that all the items within an element be listed in order. Given a sequence $\alpha = < f_1, f_2 \ldots f_m >$ (where each f_i corresponds to a frequent element in S), a sequence $\beta = < f_1', f_2' \ldots f_m' >$ ($m \leqslant n$) is called a prefix of α if and only if (1) $f_i' = f_i$ for ($i \leqslant m-1$); (2) $f_m' \subseteq f_m$; (3) all the frequent items in ($f_m - f_m'$) are orderly after those in f_m' .

Definition 2 (Suffix)[3]**.** Consider a sequence $\alpha = < f_1, f_2 \ldots f_m >$ (f_i corresponds to a frequent element in S). $\beta = < f_1', f_2' \ldots f_m' >$ ($m \leqslant n$) is the prefix of α Sequence $\gamma = < f_m'', f_{m+1} \ldots f_n >$ is the suffix of α with regards to prefix β, denoted as $\gamma = \alpha / \beta$, where $f_m'' = (f_m - f_m')^2$. Also donate $\alpha = \beta \times \gamma$. Note, if β is not a subsequence of α , the suffix of α with regards to β is empty.

Definition 3 (Projected database) [3]**.** Let α be a sequential pattern in a sequence database S. The α- projected database, denoted as $S|_\alpha$, is the collection of suffixes of sequences in S with regards to prefix α .

Definition 4 (Support count in projected database)[3]. Let α be a sequential pattern in sequence database S and β be a sequence with prefix α . The support count of β in α- projected database $S|_\alpha$, denoted as $support_{S|\alpha}(\beta)$, is the number of sequence γ in $S|_\alpha$, such that $\beta \subseteq \alpha \times \gamma$.

Here the example of how prefix span discover frequent items is given. Table 1 gives a sequence database, the process of the frequency discovery is as follows. Assume the min support equals to 2.

Sequence ID	Sequence
T10	<a1(a1a2a3)(a1a3)a4(a3a6)>
T20	<(a1a4)a3(a2a3)(a1a5)>
T30	<(a5a6)(a1a2)(a4a6)a3a2>
T40	<a5g(a1a6)a3a2a3>

Table1: Sequence Database

Step1: find frequent-1 sequential patterns. They are <a1>:4, <a2>:4, <a3>:4, <a4>:3, <a5>:3, <a6>:3. The number means the support count.

Step2: divide search space into 6 according to the result of Step 1.

Step 3: Data mining for the frequent items. The patterns can be mined by constructing the corresponding set of projected databasesand recursively. The result is shown in Table 2.

Prefix	Projected (suffix) database	Sequential patterns
<a1>	<(a1a2a3)(a1a3)a4(a3a6)>,<(_a4)a3(a2a3)(a1a5)>, <(_a2)(a4a6)a3a2>, <(_a6)a5a2a3>	<a1>,<a1a1>,<a1a2>,<a1(a2a3)>,<a1(a2a3)a1>,<a1a2a1>,<a1a2a3>,<(a1a2)>,<(a1a2)a3>,<(a1a2)a4>,<(a1a2)a6>,<(a1a2)a4a3>,<a1a3>,<a1a3a1>,<a1a3a2>,<a1a3a3>,<a1a4>,<a1a2a3>,<a1a6>
<a2>	<(_a3)(a1a3)a4(a3a6)>,<(_a3)(a1a5)>,<(a4a6)a3a2>,<a3>	<a2>,<a2a1>,<a2a3>,<(a2a3)>,<(a2a3)a1>,<a2a4>,>
<a3>	<(a1a3)a4(a3a6)><(a2a3)(a1a5)><a2><a2a3>	<a3,<a3a1>,<a3a2>,<a3a3>
<a4>	<(a3a6)><a3(a2a3)(a1a5)><(_a6)a3a2>	<a14>,<a4a2><a4a3>,<a2a3a4>
<a5>	<(_a6)(a1a2)(a4a6)a3a2><(a1a6)a3a2a3>	<a5>,<a5a1>,<a5a1a2>,<a5a1a3>,<a5a1a3a2>,<a5a2>,>
<a6>	<(a1a2)(a4a6)a3a2><a3a2a3>	<a6>,<a6a2>,<a6a2a3>,<a6a3>,<a6a3a2>

Table 2: Projected Database & Sequential Patterns

Table 2 lists all the sequential patterns in the recursive mining process of during prefix span. The specific associaiton roles can be got from the sequential patterns. This is a brief introduction of prefix span from Han Jiawei in the application of the prefix span minning vessel motion. The data details and the progress of the algorithm are modified to get a more valuable result.

3. Simulation Experiments based on Ais Data

3.1 AIS data

AIS data explain the vessel motion through multi-dimensions, including dynamic information and static information, etc. AIS data contains data location and data timestamp, and changes over time. Different types and meanings of the AIS data are listed in the following table.

Type	Meaning
MMSI	Maritime Mobile Service Identify of ships, long integer
SHIPTYPE	Vessel type, string
SHIPNAME	Vessel name, string
FLAG	Vessel flag, string
COURSE	Vessel course over ground , double
LONGTITUDE	Vessel longitude at one moment , double
LATITUDE	Vessel latitude at one moment, double
SPEED	Vessel speed , double
POSITION TIME	Accurate time of one data, time stamp

Table 3: Introduction of AIS data

About thousands of records for several vessels are chosen to conduct the experiment. All the data are cleaned before imported to the algorithm. The progress of data cleaning discards the redundancy data with same records. Some missing data items will be filled according to other records with the same MMSI.

3.2 Data Preparation

Progress for data preparation refers to the method in Ref. 9, but he only considered the location transformation. Other attributes that can describe the motion of vessels such as speed and angle should be extracted as characteristics as well. The enhancement of the data dimension can more exactly describe the motion pattern.

The method for discretizing location information is just as Ref. 9, which is also commonly used in the geographical data. The map is divided into different regions and the object motion are replaced by the transformation of the location sequence. The experiment illustration is shown in Fig 1. The region is divided into 12 $\times$10 grids with every grid possess 0.4°×0.4° in longitude and latitude. Thus the location of the vessel can be labeled by the rows and column sequence. For example, the location of a vessel in the circle can be presented as "8_6", which means the location is in the 8th row and 6th column, but it is not the final format of the vessel motion pattern. The character of speed and direction are included in the following processing period.

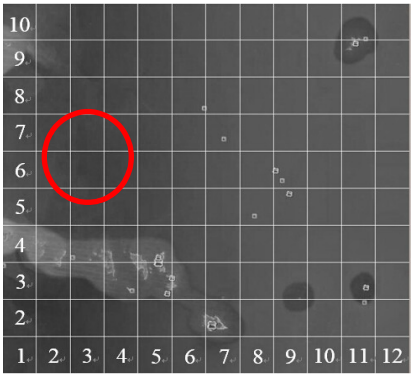

Figure 1: Grid of the Map

The method of processing the speed data is to category speed into different statuses. The result of classification is shown in Table 3.

Range of Speed (Km/h)	Result
0 ~ 3	Slow
3 ~ 14	Medium
14 ~ 23	High
23 ~ 99	Very High
Over 99	Exception

Table 4: Speed Data Processing

The method of processing the information about course over ground is to split the course into different statuses according to the 8 directions. The result of classification is shown in Table 5.

COG	Result
337.5° ~ 22.5°	N
22.5° ~ 67.5°	NE
67.5° ~ 112.5°	E
112.5° ~ 157.5°	SE
157.5° ~ 202.5°	S
202.5° ~ 247.5°	SW
247.5° ~ 292.5°	W
292.5° ~ 337.5°	NW

Table 5: COG Pocessing

3.3 AIS data in Prefix-span

In order to format AIS data into the algorithm, all the data are classified by MMSI. Every continuous trajactory is dealed as one sequence. Different trajactory are splited according to the timstamp. Records that has an interval with more than 1 hour are divided into different sequences.

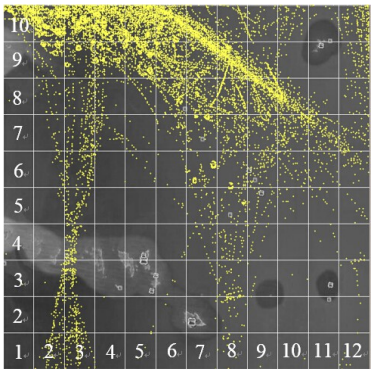

Figure 2: Trajectory Simulaiton in the Map

The data simulation with map is shown in Fig. 2. They are vessels labeled with the type "Container". The vessel trajectory with three dimensions forms the input. For example, the vessel with "8_6_Slow_S" is one element of the sequence in the algorithm. It describes the vessel moving status at that point. This vessel is in Location "8_6" with slow speed and direction to the South. Then the sequence of statues forms the sequence mined in the prefix span. The comparison of the examples in Session 2 and vessel statuses here is shown in Table 6 below.

Sequence ID	Sequence	MMSI	**Sequence**
T10	<a(abc)(ac)d(cf)>	1XXX	<(10_7_Slow_SE, 9_8_ Medium_SE, ...)>
T20	<(ad)c(bc)(ae)>	2XXX	<(4_3_Slow_S, 4_3_Slow_S...)>
T30	<(ef)(ab)(df)cb>	3XXX	<(6_5_ Medium_SE, 5_6_ Medium_SE), (7_4_Slow_SE...)>
T40	<eg(af)cbc>	4XXX	<(2_2_Medium_S,1_2_ Medium_S...)>
T10	<a(abc)(ac)d(cf)>	5XXX	<1_8_Slow_N,1_8_Slow_N...>

Table 6: Input Data for Example and Vessel Statuses

In this paper, the minmun support count is supported to 10. The processing progress of the algorithm in the following:

Input: A status sequence database S, and the minimum support threshold.

Output: The complete set of sequential patterns.

Method: *PrexSpan* (<>; 0; S).

Subroutine: PrexSpan(αl;S|α) is described below.

Stage 1: Scan S|αonce find each frequent item, b, such that

(1) b can be assembled to the last element of $\alpha$to form a sequential pattern; or

(2) < b > can be appended to $\alpha$to form a sequential pattern

Stage 2: for each frequent item b, append it to α to form a sequential pattern α', and output α'.

Stage 3: for each α', construct $\alpha' - projected$ database S| α', and call PrefixSpan(α)

After processing, the frequence items can be discovered. An example is shown in Table 7. It is the base for further analysis and anomaly detection.

Prefix	Frequence items
<10_7_Slow_SE>	<10_7_Slow_SE>,<10_7_Slow_SE, 10_7_Slow_SE> <10_7_Slow_SE,9_8_Medium_SE>, <10_7_Slow_SE,9_8_Medium_SE, 8_9_Medium_SE> .....

Table 7: Example of Frequence Items

3.4 Get Crucial Status

Prefix span can discover frequent items, which is represented as subsequence. To find out specific asscociation for different items, further calculation needs to be conducted.

Step 1: get every prefix status.

Step 2: for every status, get every frequent items

Step 3: count the frequency statues of the frequent sequence and ensure the operation happens in the same trajectory.

Step 4: the max n frenqueny status is the n highestly strongly associated.

3.5 Result analysis

3.5.1 Result for Maritime Surveillance

The result can be used for the maritime surveillance. The Surveillance can be conducted in two aspects.

The first aspect is the suvillance for the crucial statuses. For example, the "10_7_Slow_SE" and the "9_8_Medium_SE" are both crucial statues. The improtance for surveillance is to find out the vessels with the direction not SE or the speed not slow in the position 10_7 and find out the vessels with direction not SE or speed not medium in the position 9_8.

The second aspect is the surveillance for trajectory. The frequent items illustrates the busy trajectory and it features such as speed and direction in the maritime traffic. It gives strongly support for the management of intelligent maritime traffic.

3.5.2 Result for Anomly Detection

Result for the anomy detection mainly depends on the frequent item list. If the vessle is witnessed in the statuses of prefix in Table 7, which is totally over 1,000 statues. Then its motion can be described according to the frequent items. If the motion is not in accordance with the expected one. Then the anomly probably happens. For the algorithem deals with three dimensions, the anomly can occur in the speed, direciton or the location.

4. Conclusion

This paper introduces the prefix span algorithem to process AIS data. This processing progress can reduce the space and time scale, which is a huge improvement. The data dimension is extended in the progress. In this paper, more information is dealed with. The result can be used for maritime surveillance and traffic management, and it can strongly support the anomly detection for vessel motion. The cluster for trajectory and data fusion of multiple sources can be introduced in the preprocessing period. The overall change of the region can be analyzed in the anomly detetion, as the crucial point in the future work.

References

[1] J. Y. Kang, S. Y. Hwan,*Mining Spatio-Temporal Patterns in Trajectory Data.* JIPs, 2010. **6**(4): p. 521-536.

[2] K. B. James, A. L. Scott and W. L. Micheal. *Automated anomaly detection processor,* in *AeroSense 2002*. 2002. International Society for Optics and Photonics.

[3] J. Pei, J. Han, et al., *Mining sequential patterns by pattern-growth: The prefixspan approach.* Knowledge and Data Engineering, IEEE Transactions on, 2004. **16**(11): p. 1424-1440.

[4] N. Yusof, et al., *Mining Frequent Spatio-Temporal Patterns in Wind Speed and Direction*, in *Connecting a Digital Europe Through Location and Place*. 2014, Springer. p. 143-161.

[5] J. J. C.Ying,W. C. Lee ,T. C. Weng, et al. *Semantic trajectory mining for location prediction.*Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2011: 34-43.

[6] R. Assam, T. Seidl. *TMC-pattern : holistic trajectory extraction, modeling and mining.* Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data. ACM, 2012: 71-80.

[7] X. Z. Wang, T. Chen, Q. Lu, L. C. Sun, et al. *Software Behavior Trustworthiness Detection Based on PrefixSpan Method.* Proceedings of the 2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control. IEEE Computer Society, 2011: 508-511.

[8] A. Rakesh, I. Tomasz, S. Arun. *Mining association rules between sets of items in large databases.* ACM SIGMOD Record. ACM, 1993, 22(2): 207-216.

[9] F. X. Zhu. *Mining ship spatial trajectory patterns from AIS database for maritime surveillance.* Emergency Management and Management Sciences (ICEMMS), 2011 2nd IEEE International Conference on. IEEE, 2011: 772-775.

[10] F. Deng, S. T. Guo, Y. Deng, H. Y. Chu, Q. M. Zhu, and F. C. Sun. *Vessel track information mining using AIS data.* 2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI),2014.

[11] Y. C. Zheng, F. Deng, Q. M. Zhu and Y. Deng. *Cloud storage and search or mass spatio-temporal data through Proxmox VE and Elasticsearch cluster.* 201EEE 3rd International Conference on Cloud Computing and Intelligence Systems, 2014.