TRANSNAV International Journal on Marine Navigation and Safety of Sea Transportation
Volume 6 Number 2 June 2012

Applied Research of Route Similarity Analysis Based on Association Rules

Z. Xiang, R. Liu, Q. Hu & C. Shi
Merchant Marine College, Shanghai Maritime University, China

ABSTRACT: In recent years, with the development of information technology, businesses have accumulated a lot of useful historical data, as the shipping industry does. These data can be found deposited a large number of "knowledge", for example, Shipping records for historical information, Ship-Port relations information, Ship-ship relations information, Port & shipping route relations, Shipping route information. It can provide intellectual support to shipping informatization development.
Association rules in data mining technology is one of important technologies. The technology, based on statistical methods, can mine the associated and implied "knowledge" from data warehouse ,which has a large number of accumulated data. Apart from this, the technology can also play an important role in the prediction. In this paper, based on FP-growth algorithm, we improve it forming Relevent ships routes.
From the prevalent perspective of data mining, deal with the corresponding vessels' dynamic information, acquired from the AIS, such as data collection, data statistics. On this basis, get the ship-port relation and ship-ship relation after a certain level of data analysis, processing, handing. Furthermore, this paper use the numerous historical ship-port relation and ship-ship relation to build a mathematical model on the ship-port and ship-ship relation. And use the improved association algorithm, FP-growth algorithm, to acquire the strong association rules between ship-port and ship-ship, and eventually mine the similarity of the ship route.
Main points of this paper as follows:
Collect ,count and check the data, which is from ship dynamic information;
Establish the mathematical model between ship-port and ship-ship relation;
Improve the algorithm;
Analyse the similarity of ship route more accurately using the improved algorithm.

1 INTRODUCTION

In recent years, data mining has caused great concern to the industry. The so-called mining technology is to present large amounts of data by digging into useful information and knowledge, which applies to business management, production control, market analysis, engineering design and scientific exploration and other fields. This is the data mining technology on information processing in navigation, the use of data mining association rules FP-TREE algorithm, mining the large number of ships AIS information, access to the ship's route similarity analysis.

1.1 *FP-TREE algorithm [1]*

Data mining technology is the result of the development of information technology, since the 60 years since the last century, databases, and information technology has systematically evolved from the original document processing to complex and powerful database system, and now the technology is quite mature. Data mining is the definitions of raw data from a large number of extracted or “dig out" the useful information into knowledge.

In the knowledge model of data mining, association rules model is the more important one. The concept of association rules by the Agrawal, Imielinski, Swami suggested that the data in a simple but very useful rule. Association rules models are descriptive models, association rules discovery algorithm is unsupervised learning.

FP-TREE [1][2][3] is one of association rule mining algorithm, which uses divide and conquer strategy, after the first pass scan, the frequency of the database into a set of compressed frequent pattern tree (FP-TREE), while retaining the associated information, then FP-TREE library to differentiate into a number of conditions, the length of each library, and a frequency of 1 set of related libraries of these conditions were re-excavation.

1.2 *AIS information [4]*

Automatic Identification System (AIS) is the current advanced ship-aided navigation equipment; the International Maritime Organization has adopted the mandatory installation of AIS requirements. AIS can automatically send the ship a static continuous, dynamic and voyage information, security, short message, but can also automatically receive the information sent around the ship, and exchange information with the coast station.

AIS information includes static information such as name, call sign, MMSI, ship type, ship size and other information, and dynamic information such as vessel position, ground speed, navigational status, draft, destination, estimated time of arrival and other information, but also with the voyage information and security-related text messages.

1.3 *AIS information on data mining [5][6]*

In this paper, using association rules such as FP-TREE algorithms, we deal with the corresponding vessels' dynamic information, such as estimated time of arrival, port of destination and departure information, acquired from the AIS, such as data collection, data statistics. On this basis, get the ship-port relation and ship-ship relation after a certain level of data analysis, processing, handing to build a mathematical model on the ship-port and ship-ship relation to acquire the strong association rules between ship-port and ship-ship, and eventually mine the similarity of the ship route.

2 THE PROPOSAL OF THIS PAPER

At first, we used to provide users dynamic-related information services with AIS ship data, such as real-time latitude and longitude of the ship, speed, arrival time, etc. Also including the port border querying. But we are lack of a more intelligent data mining reports service, for example, previous leaves harbor of the ships, the ships' similar route. These are the users concerned, and it is very useful.

Therefore, in order to provide such services, we need to mine existing data to find out the regular pattern between ships and their anchored port and also we need to find out ships' similar routes(the similar route among ships).The conclusion we mine from AIS data can provide information on the effect of enhanced information services . We collected ships infomation through the www.manyships.com dynamic data, and similarly, the Chinese port border has also been part of we collect. It is easy for us to find the ship's ports and similar routes and provide us a lot of supports for data ming.

The most important task of this paper is: data mining, to identify routes of ships similar to the (first find was anchored in this port).

3 PREPARATION FOR DATA MING

3.1 *Data collection*

This article uses real-time AIS data from www.manyships.com. AIS database established a table for each ships, with the AIS data updating continuously the database update this table as table 1.

The number of table is Real-time. So, the first step we need to do is to scheduler an interval job to collect data using database snapshots[4]. With this methods,we can collect one week or one month even more AIS data[5] . To say that, for accurcy results,job's interval can not be too long.

3.2 *Data processing and Summary*

After collecting the data,we must immediately calculate whether the ship is in the ports and also calculate its last port .

Finally we proceed the AIS data for each ships. These data included the ship's number(mmsi), name, port of arrival, arrival times, the last port, the last departure time, callsign, type, so as follow. These data be prepared for data mining as table 2.

Each port has its own id, so 'fid' in the picture means the port's id , 'prevfid' means the last port's id. With these data we can be ming AIS data.

Table 1. Original datasets snapshot

mmsi	speed	lon	lat	course	heading	updatetime
565101000	.100000001490116(	7290.172800000000	2260.931200000000	275.200000000000(	177.000000000000(	2010-12-04 08:10:07
413552790	8.600000381469730	7313.721600000000	2427.645000000000	219.000000000000(	238.000000000000(	2010-12-03 07:08:02
413427340	8.399999618530269	7317.120000000000	1831.393600000000	264.700000000000(	17.0000000000000(	2011-01-07 07:24:19
477383000	.000000000000000(	7151.783200000000	1933.761000000000	317.200000000000(	306.000000000000(	2011-01-07 07:24:21
412206710	9.699999809265140	7186.139200000000	1551.126600000000	208.100000000000(	-1.00000000000000	2010-12-04 08:12:18
412047720	.000000000000000(	7295.708000000000	1882.021200000000	204.100000000000(	-1.00000000000000	2011-01-07 07:24:58
412047210	.200000002980232(	7325.068800000000	1864.132400000000	123.100000000000(	292.000000000000(	2011-01-07 07:21:50

Table 2. Datasets snapshot after procceed

mmsi	leavetime	fid	fid_name	prevfid	prevn...	prevleavetime	name	callsign	type
412051550	2010-08-30 12:46:59	1	<MEMO>	47	<MEMO>	2010-08-28 08:28:08	YONG CHI	BRYU	货轮
412053050	2010-08-26 14:29:23	1	<MEMO>	5	<MEMO>	2010-08-21 20:29:01	WAN QING SHA	BSPN	疏浚或水下作业船
412064000	2010-08-26 23:29:05	1	<MEMO>	207	<MEMO>	2010-08-23 05:49:23	AN GUANG JIANG	BOAU	货轮
412070000	2010-08-21 23:27:06	1	<MEMO>	42	<MEMO>	2010-08-20 05:55:30	XING HE	BOLO	货轮
412070630	2010-08-28 15:08:07	1	<MEMO>	201	<MEMO>	2010-08-27 08:29:45	HANG CE 501	未知	其他类型船
412081690	2010-08-27 21:13:05	1	<MEMO>	204	<MEMO>	2010-08-23 20:33:19	CHANG HANG SHA	BUVU	货轮
412187000	2010-08-24 12:23:27	1	<MEMO>	4	<MEMO>	2010-08-23 10:07:13	XIN BIN CHENG	BVFB5	货轮
412205920	2010-08-31 22:52:02	1	<MEMO>	5	<MEMO>	2010-08-27 04:29:26	XIN YUN FENG	BAOZ	货轮
412207640	2010-08-22 23:27:31	1	<MEMO>	202	<MEMO>	2010-08-22 04:21:59	GANG TONG HAI 9	BANA	未知
412222000	2010-08-28 01:54:29	1	<MEMO>	2	<MEMO>	2010-08-19 20:57:58	TUO HAI	BOHF	货轮
412258000	2010-08-29 01:17:33	1	<MEMO>	31	<MEMO>	2010-08-25 05:05:08	YING CHUN HAI	BIAE3	未知
412272000	2010-08-29 23:58:47	1	<MEMO>	207	<MEMO>	2010-08-20 18:13:01	FENG AN SHAN	BOST	货轮
412351000	2010-08-30 14:54:22	1	<MEMO>	5	<MEMO>	2010-08-25 07:30:10	ZI YUN FENG	BHVV	未知

4 MINING THE SIMILAR PATH AND THE RELEVENT PORT

In order to retrieve the relative path between the ship, we use association rule discovery methods to determine the correlation between the ports and the ships.

The central idea of this method is: Calculate the ship reaches port summary statistics. Use summary information as a transaction set; The port each ship reached (or leave) is the association rules item. Each transaction is a single arrangement of the ship reaches port. Aim of the algorithm is digging out the relevant port, and getting the ships' similar paths.

4.1 *Algorithm is defined as follows[1][3]:*

(1). $I = \{I_1, I_2 ..... I_m\}$ is item set.

(2). Let D transaction sets, each transaction T is $T \subseteq I$.

(3). Each transaction has a unique identifier TID. In addition, add a field to indicate the transaction from the ship (the number of ship).

For an example of Transaction sets is as table 3:

Table 3 example of Transaction sets

TID	PORTID	Mmsi
1	200 , 47 , 48, 203 , 56 , 207 , 88 , 1	412402820
2	47 , 21, 48, 200, 66 , 71	412403000
3	21, 200 , 42 , 59 , 88 , 1	412403660
4	21, 48 , 49 , 88 , 71, 6	412410010
5	47 ,200 , 48 , 62 , 66 , 51 , 71, 1	412429760

Each port has an unique identity number ,so as ships. In this paper, Example uses the portid and mmsi to instead of the specific ports and ships.

4.2 *Algorithm Description:*

We intend to make a FP-tree for the transaction sets, the tree node is a transaction set, namely the port (except root node). Edge of the FP-tree is ships' collection, on behalf of the ship go across the two ports. We use 'Mmsi' instead of ship's name.(mmsi is the unique identity of the ships in database).

The structure of Node data as table 4 :

Table 4 The structure of Node

Parent	Portid	Support	Child

The structure of Node data contains the number of ports, its support count, its child node, its parent node.

The structure of Edge data is as table 5:

Table 5 The structure of Edge

PortID1	Mmsi[]	mmsi	Support[]	PortID2

Edge data structure as shown above, Mmsi [] expressed as mmsi array, a collection of both ships Mmsi number. Similarly, the corresponding number of ships mmsi also have the corresponding support count. PortID1 and PortID2 mean this edge belongs to which two port nodes.

4.3 *Steps of the algorithm*

According to the minimum support, we use the transaction sets generated a FP-tree by FP-growth algorithm.

1 [1][2] scan transaction set D, in order to acquire all the frequent items contained in D ,we named the collection of these items F,and also caculate their respective support. Frequent items in de-

scending order according to their support, the results recorded as L(table 6):

2 [1][2] Create the root of FP-tree T, marks "null"; Then, for each transaction TID following:
According to the order of L to select and sort the frequent items TID. Let the sorted frequent item list as [x|P], x is the first frequent item, and P is the remaining frequent items; Then call INSERT_TREE([x|P],T).The INSERT_TREE([x|P], T) follows the process of implementation: If T have their children named N and meet the N. Portid = x. Portid, increase the N's Support 1; else create a new node N, its count is set to 1, link to its parent node T and through the node chain structure link to the node with the same Portid. If P is not empty, recursively call the INSERT_TREE (P, N).
In addition, whether the addition of new nodes, each edge must be coupled with corresponding Mmsi TID number. if repeated, Mmsi Support count increased by 1.
Re-scanned in the transaction set, a complete with side information on FP-tree built.
Using the example 3.1 , let the minimum support MIN_SUPPORT = 3, the 1-frequent itemsets is as follows:

1 - frequent item sets as table 6 :

Table 6 1 - frequent

item	count
200	4
48	4
47	3
21	3
88	3
71	3
1	3

And we also acquire the FP-tree as figure 1.

The information of edge as table 7:

Table 7 information of edge after calculated

Edge ID	PortID1	Mmsi[]	mmsiSupport[]	PortID2
1	200	412402820, 412403000, 412429760	4,4,2	48
2	48	412402820, 412403000, 412429760	3,3,1	47
3	47	412403000	3	21
4	21	412403000	1	71
5	47	412402820	1	88
6	88	412402820	1	1
7	47	412403000	1	71
8	71	412402820	1	1
9	200	412403660	1	21
10	21	412403660	1	88
11	88	412403660	1	1
12	48	412410010	1	21
13	21	412410010	1	88
14	88	412410010	1	71

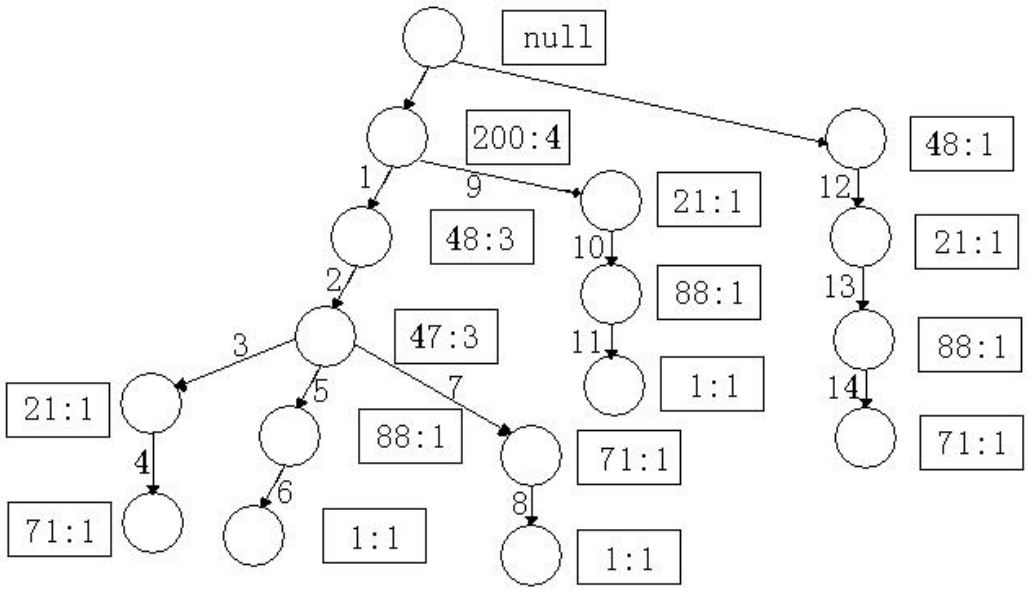

Figure 1 generated tree

4.4 *Mining Rules*

We has generated a FP - tree, according to frequent 1 - itemsets, we can mine the frequent pattern as table 8:

Table 8 frequent pattern of Ports

id	frequent pattern of Ports
1	200-1
2	200-47
3	48-47
4	200-48-47
5	200-48
6	48-71

Meanwhile, according to the minimum support (min_support) branch cut, we also can get the ship-related items as table 9:

Table 9 frequent pattern of ships

Id	frequent pattern of ships
1	412402820-412403000

This result shows that a total of 6 groups of ports are associated. What's more,the MMSI number which is 412402820 and 412403000 of the two ships has some similar shipping routes.

Thus, by calculating the ship - port summary information,we not only acquire the Correlation of the ports but also the the Correlation of the ships' routes.

5 CONCLUSIONS AND SUMMARY

In this paper,based on the AIS data analysis and processing, we use improved FP-growth algorithm,the algorithm of data mining association rules to analyze the ship arrival information, and build FP-tree, by scanning the transaction sets, creating FP-tree root sub-set of structural conditions of frequent library, and mining the sub-set to get frenquent pattern.Through all these methods,we aggregate and mine the information of ship and port, then can find out the relevant port and ships with the similar routes.

Use of the collected AIS data which is more than one week to run above algorithm method ,we get 42 groups of related ports and 105 ships.(if you want to get more accurate results, you need to capture more of the AIS data) This shows the feasibility of the algorithm. In addition, compared to other association rule algorithm, this algorithm only need to scan the database twice.

6 DEFICIENCY AND OUTLOOK

In this paper,we use improved FP-tree Algorithm to dig out the similar route and relevant port,but deficiency as follow:

1. FP-tree need lots of Server's memory.
2. how to set minimum support and confidence is not in-depth reserch.

Thus,our further research will focus on improving the efficiency and application. In addition, data mining application in the navigation area is also the direction of future research

REFERENCE

[1] MA Xu-hui, Zhang A-hong. Association Rules Generated By the FP Tree Depth-First Algorithm. Computer Knowledge and Technology, 2010, 13: 058

[2] HUI Liang, QIAN Xue-zhong. Non-check mining algorithm of maximum frequent patterns in association rules based on FP-tree. Journal of Computer Applications, 2010, 07: 064

[3] YAN Wei, BAI Wen-yang, ZHANG Yan. Hiding Association Rules with FP-Tree Based Transaction Dataset Reconstrucion. Department of Computer Science and Technology, 2008

[4] JI Xian-biao, SHAO Zhe-ping, PAN Jia-cai, et al. Development of distributed data collection system of AIS information and key techniques[J].Journal of Shanghai Maritime University, 2007, 28(3): 28-31

[5] Lee C Y. An algorithm for path connections and its applications. IRE Trans Electronic Computers, 1961, EC-10: 346-365

[6] ZHANG Wen, TANG Xi-jin, YOSHIDA Taketoshi. AIS: An approach to Web information processing based on Web text mining. Systems Engineering-Theory & Practice, 2010, 01:015