Mining Association Rules from Stream Data Based on the Dynamic Support

Jia Luo[1], Shihe Chen[1], Fengping Pan[1], Yaqin Zhu[1], Le Wu[1], Yaqi Sun[2] and Chunkai Zhang[2*]
[1]Electric Power Research Institute of Guangdong Power Grid Co., Ltd. China
[2]Department of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China

***Abstract*—The Stream data exists in the field of industrial production, life activities, business transactions, and other fields. It is closely related to people's life, production and so on. This paper proposes inter-transaction association rules mining method based on dynamic support threshold. Inter-transaction association rules refer to the association rules between different time periods. This paper firstly uses the sliding window to limit stream data, then do preprocessing on stream data. In the process of pretreatment using linearization method fitting to raw data and it reduce the amount of data at the same time, and finally at the end of the preprocessing, generating large transaction grouping method of inter transaction association rules is proposed in this paper. This paper uses conceptual data attenuation, thereby reducing the influence of old data to the mining result. Due to artificial setting minimum support threshold may bring many problems, so this paper presents a method for searching minimum support threshold.**

Keywords-stream data; association rules; inter transaction; support threshold.

I. INTRODUCTION

The industry data, the network data, the business data and so on are generated in the form of data flow. The characteristics of data stream itself so that it presents a unique challenge for mining algorithm. A inter transaction association rule[1] describes the relation-ship between different transactions, and it can analyze the relationship between different time and different time series. These association rule have important significance for us to predict the occurrence of events.

In this paper, we have improved the cross transaction association rules mining from the following three aspects:

- Data across the inter transaction set method is proposed to generate large packet services, by pruning the big data.
- Using the concept of data attenuation[2] to reduce the impact of historical data on the current mining results.
- A method for searching the minimum support threshold is proposed, which can solve the problem that the minimum support threshold[3] may be brought to a lot of problems.

After the improved algorithm is called: ITF-tree algorithm.

The second part introduces the basic concept of this method. The third part is the preprocessing of time series, the fourth and fifth part is the mining algorithm of cross transaction association rules. The sixth part describes the experimental results, the seventh part is the summary.

II. BASIC CONCEPTS

A. Sliding Window

Limiting the convection data using sliding windows[4]. The data in the sliding window can be described like: If the length of the window is n, data in a window at a time point T_i is:

$$Di = \frac{\{S0, S1, ...Si\} i \le n-1}{\{Si, Si+1, ...Sn+i-1\} i \ge n-1} \quad (1)$$

where D_i is the data set of the entire sliding window, S_i only indicates the data values at the time point i.

In normal circumstances, a sliding window is composed of the same fundamental window of several sizes. Sliding window $SW = \{FW_1, \ldots FW_i \ldots FW_n\} (1<i<n)$, FW_i is the ith fundamental window in sliding window. The following figure represents a fundamental window.

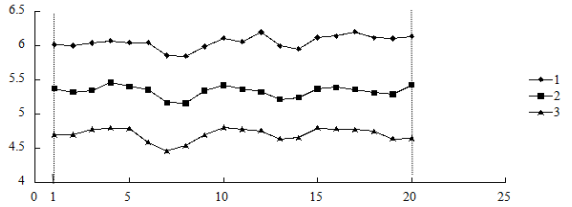

FIGURE I. THE FUNDAMENTAL WINDOW OF TIME LENGTH 20

III. DATA PREPROCESSING

Data sets can be transformed into inter-transaction data set by three steps.

A. Linearization and Generating Meta Model

Linear regression method is used to generate fitting line segments. In the linear regression, the least square method[5] is used to obtain the least error of fitting line segment.

The line segment is the original segment after the linearization, then it is to divide the time series and generate the meta model[6]. The results of different data flow lines are shown in Figure 2.

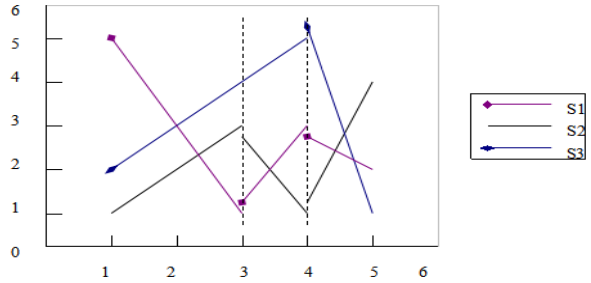

FIGURE II. THE RESULTS OF DIFFERENT DATA STREAMS ARE LINEARIZED

In order to facilitate, that is only one segment of the same data stream in a time period. We split the data stream *S*3 at the time point of 3, and then In the time period 1 to 5 there are three segments are: 1-3,3-4,4-5,the same as *S1* and *S*2.

For clustering segment, Firstly, the representation of the similarity between segments is to be performed. It can be said to be a measure of the distance of the line. The smaller the distance, the smaller the similarity is. This paper proposed a method of measuring the similarity of two different segments based on cosine similarity[7]. Assume that there are two line segments, respectively, *i* and *j*, The formula (2) is expressed as follows:

$$SIM_{i,j} = \omega_1 sim_{Dir} + \omega_2 sim_{Len} + \omega_3 sim_{y0} \quad (\omega_1 + \omega_2 + \omega_3 = 1) \tag{2}$$

sim_{Dir},sim_{Len},sim_{y0} espectively express two mode on the slope of the similarity, Similarity of X axis length, the similarity of the starting point Y0 height. SIM_i represents the similarity of line i and j. w1，w2，w3 represent three aspects of weights.

So we can define the above three parameters, Respectively, using the formula (3), (4), (5) express.

$$Sim_{Dir} = \cos(\tan^{-1}(\left|\frac{i_a - j_a}{1 + i_a j_a}\right|)) \tag{3}$$

$$sim_{y0} = \begin{cases} \dfrac{1}{|i_{y0} - j_{y0}|} & |i_{y0} - j_{y0}| > 1 \\ 1 - |i_{y0} - j_{y0}| & |i_{y0} - j_{y0}| \le 1 \end{cases} \tag{4}$$

$$sim_{Len} = \begin{cases} \dfrac{1}{|i_L - j_L|} & |i_L - j_L| > 1 \\ 1 - |i_L - j_L| & |i_L - j_L| \le 1 \end{cases} \tag{5}$$

In which i_a, i_{y0}, i_L stands for the slope of the line *I*, height y_0 of the starting point on the Y axis, the length *L* of the X axis, respectively. With the similarity, we can use the general clustering algorithm to cluster. In this paper, we have used the most effective K-means method.

B. *Generating Transaction set*

By linearization and sign, we have generated the original data set by different symbols representing the pattern data set. Next, we must convert the time series data set into a transactional data set.

We generate the transaction data set each item marked the flow number of the item, So in the transaction set, each transaction is composed of two parts:(Numerical, value).In the program we have connected with the numerical flow numbers underline "_".

IV. INTER TRANSACTION FREQUENCY TREE ITF-TREE

A. *ITF-Tree Data Structure Overview*

ITF-tree mainly consists of two parts: the head node list and tree structure.

The head node list includes the following:

Item_name: Item name.This item indicates which of the items is stored in this table.

Support_count: Support count. It indicates the number of times that item_name appears in the data set.

ITF-tree_pointer: It points to the node of a tree with a item_name name in the tree.

ITF-tree tree is the main body of the data structure of the algorithm. The structure is:

Item_name: Item name which of the items is stored in this node.

Count: Count of support for the node on all paths of the node.

Brother_ITF-tree_pointer: It points to a pointer to the node with the node item_name.

Parent_ITF-tree_pointer: It points to the father node of the node.

B. *Construction and Maintenance of ITF-tree*

Suppose we now get the transaction data set is shown in Table 1.

TABLE I. TRANSACTION DATA SET

LTID	ItemSets
LT100	S1_A_0,S2_G_0,S3_C_0,S1_A_1,S2_B_1,S3_B_1,S1_B_2,S2_D_2,S3_B_2
LT200	S1_A_0,S2_B_0,S3_B_0,S1_B_1,S2_D_1,S3_B_1,S1_A_2,S2_A_2,S3_B_2
LT300	S1_B_0,S2_D_0,S3_B_0,S1_A_1,S2_A_1,S3_B_1,S1_A_2,S2_B_2,S3_C_2
LT400	S1_A_0,S2_A_0,S3_B_0,S1_A_1,S2_B_1,S3_C_1,S1_C_2,S2_A_2,S3_C_2

We can get a sequence of counting the individual items:(S1_A_0, 3), (S1_A_1, 3), (S3_B_0, 3), (S3_B_1, 3), (S2_B_1, 2), (S1_A_2, 2), (S2_A_2, 2), (S3_B_2, 2), (S3_C_2, 2), the rest of the individual items are 1.Suppose the minimum support threshold is 2,the transaction of non-frequent itemsets removed, reducing the amount of data. Get the following set of things:

TABLE II. DATA SETS FOR AN ORDERLY TRANSACTION

LTID	Items
LT100	S1_A_0,S1_A_1,S3_B_1,S2_B_1,S3_B_2
LT200	S1_A_0,S3_B_0,S3_B_1,S1_A_2,S2_A_2, S3_B_2
LT300	S1_A_1,S3_B_0,S3_B_1,S1_A_2,S3_C_2
LT400	S1_A_0,S1_A_1,S3_B_0,S2_B_1,S2_A_2,S3_C_2

In Figure 3 below,blue arrow indicates that a node from the list node with the same item_name in the tree,the black arrow indicates the parent node to the child node,red arrow connecting the two nodes in the tree.These two nodes have the same item_name. A green arrow indicates pointing to the parent node.

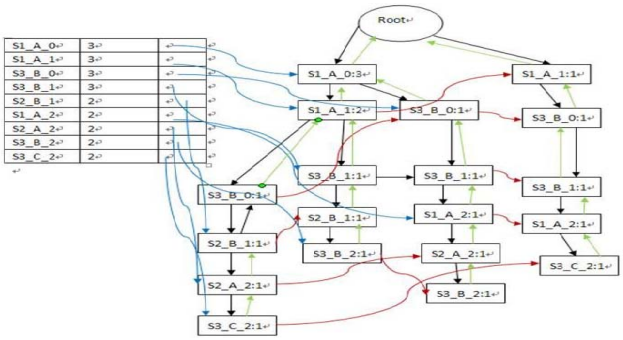

FIGURE III. ITF-TREE STRUCTURE

V. SELECTION OF DYNAMIC SUPPORT

In general association rule mining method, we need to artificially set a subjective sense of minimum support threshold min_sup.

A new evaluation criterion is proposed in this paper: Suppose the current minimum support threshold is min_sup, the number of frequent itemsets excavated under this threshold is counter, The evaluation index can be written as a formula (6)

$$f(\min\_\sup) = \min\_\sup * counter \tag{6}$$

In practical application we set the min_sup between 0 and 1,The formula can be rewritten as a formula (7).

$$f(\min\_\sup) = \min\_\sup * \ln(counter) \tag{7}$$

We have carried out the statistics on multiple data sets.It is found that the number of frequent itemsets generated under this support is approximately to figure 4:

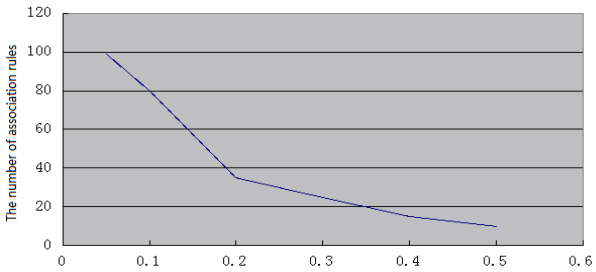

FIGURE IV. MINIMUM SUPPORT THRESHOLD AND ASSOCIATION RULES

In this case, the negative high order is used to fit the curve:

$$y = a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \frac{a_3}{x^3} \tag{8}$$

which X expressed support threshold, Y represents the number of frequent itemsets generated under this threshold. Make the following transformation:

$$y_1 = x^{-1} \tag{9}$$

$$y_2 = x^{-2} \tag{10}$$

$$y_3 = x^{-3} \tag{11}$$

Therefore, put the formula (9) (10) (11) into the formula (8)，we can get the following formula:

$$y = a_0 + a_1 y_1 + a_2 y_2 + a_3 y_3 \tag{12}$$

Let y equal to 1 can be obtained as follows:

$$y = \sum_{i=0}^{3} a_i y_i \tag{13}$$

Then we get the following formula:

$$f(min\_sup) = \ln(a_0 + \frac{a_1}{min\_sup} + \frac{a_2}{min\_sup^2} + \frac{a_3}{min\_sup^3}) * min\_sup \tag{14}$$

Maximum values between intervals (0, 1),the formula (14) is about the min_sup curve,So we can use the three points of the introduction of random factors to find the extreme value of the method.

VI. EXPERIMENTAL RESULTS

In order to test the performance of ITF-tree in large data sets, so compare the performance of the algorithm with the traditional Apriori and FP-growth on the stock data set. The data from Biaopu Yonghua website. From July 1997 to December 2008, including Shanghai and Shenzhen two stock 5 minutes K-line data. This paper take the Shanghai Composite Index——SH1A0001, A stock index——SH1A0002，B stock index——SH1A0003, composite index——SH1B0006 consisting of Stock A data sets. Take the Shanghai Industrial Index——SH1B0001, business index——SH1B0002, Real estate index——SH1B0003, utilities index——SH1B0004 consisting of Stock B data sets. The data set Stock A and Stock B fusion together for Stock C. Data set up to a million levels. SH1A0001 fitting threshold was set to 2.2, the fitting of SH1A0002 was set to 2.6, the fitting of SH1A0003 was 0.095,

SH1B0006 was 2.5, SH1B0001 was 1.6, SH1B0002 was 2.1, SH1B0004 was 4.1 and SH1B0005 was 3.4. The window size is 500, the window size is 10, the support threshold is 0.1, and the confidence threshold is 0.5. FP-growth and Apriori and ITF-tree running time in the data set that have been preprocessed are shown in figure 5.

Performance of the algorithm on the stock data set.

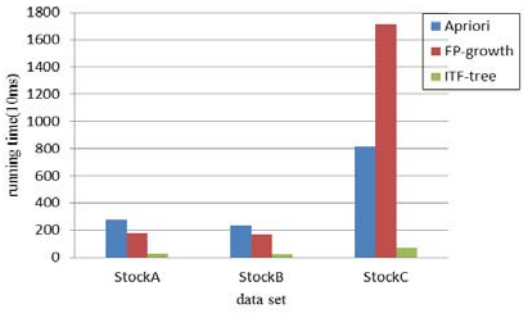

FIGURE V. RUNNING TIME UNDER DIFFERENT DATA SETS

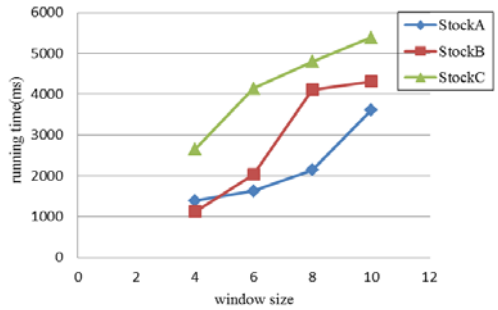

FIGURE VI. ALGORITHM RUNNING TIME OF THE WHOLE PROCESS

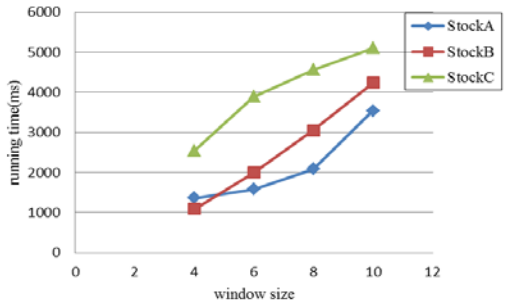

FIGURE VII. PREPROCESSING TIME

VII. CONCLUSION AND SUMMARY

In this paper, the sliding window method is used to limit the data, and then through the linear processing, similarity clustering etc pretreatment. Finally, the association rule is the association rule of the transaction. In this paper, we propose a method to generate the transaction data set based on the previous research. ITF-tree algorithm is proposed in this paper based on FP-growth algorithm. ITF-tree is a summary of data mining. In order to reduce the impact of historical data on the current data mining, the concept of data attenuation is introduced. This paper presents a new method for evaluating the degree of support based on the results of previous studies. And on this basis, a better support threshold is obtained. Experimental results show that the algorithm is very effective.

REFERENCES

[1] Anthony K.H. Tung, Hong Jun Lu, Jiawei Han, et al. Breaking the Barrier of Transactions: Mining Inter-Transaction Association Rules[J]. IEEE Transactions on Knowledge & Data Engineering, 2003.

[2] Sayal M. Detecting time correlations in time-series data streams[J]. Hewlett-Packard company, 2004.

[3] Grahne G, Zhu J. Fast algorithms for frequent itemset mining using FP-trees[C]//IEEE Transactions on Knowledge and Data Engineering. 2005: 1347-1362.

[4] Han J, Cheng H, Xin D, et al. Frequent pattern mining: current status and future directions.[J]. Data Mining & Knowledge Discovery, 2007, 15(1): 55-86.

[5] Agrawal R, Imielinski T, Swami A N. Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Conference[J]. Acm Sigmod Record, 1993, 22:207-216.

[6] Chai S, Yang J, Cheng Y. The research of improved apriori algorithm for mining association rules[C]//Service Systems and Service Management, 2007 International Conference on. IEEE, 2008: 513-516.

[7] Sun Fan. Research on the association rule mining technology [D], Liaoning Normal University,2012.