Acta Technica *62* (2017), No. 4A, 521–528 © 2017 Institute of Thermomechanics CAS, v.v.i.

Application of improved time series apriori algorithm in data mining of association rules based on temporal constraint

Chunxia WANG[2]

Abstract. The basic idea of Apriori algorithm is to find all the frequent sets in the transaction, and the frequent need of these frequent sets is greater than or equal to the minimum support degree of the set. Then the paper describes the working principle of the traditional Apriori algorithm, and points out the existing problems. To solve these problems, this paper proposes an improved Apriori algorithm for frequent item set time series. In this paper, we analyze the methods and procedures of time series and time sequence association rules mining. The paper presents combined application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint. The last experiment shows that the improved Apriori algorithm is better than the traditional method in the storage space.

Key words. Time series, apriori algorithm, frequent item, data mining, association rule, temporal constraint.

1. Introduction

In temporal data mining, temporal data can be divided into time series, sequence of events and event sequence. The time series is a narrow sense of time series, and its sequence elements are numerical. With many algorithms of association rules and algorithms, more and more people are involved in the research of association rule algorithm [1].

Data mining is a very important and promising new field in database and related fields. At present, the research on data mining is mainly focused on classification, clustering, association rule mining, sequential pattern discovery, and anomaly and

[1] Acknowledgement - This paper is supported by Soft Science Research Project of Henan Province (No. 132400410979, 132400410905) and Key research projects of Henan colleges and Universities(18A520008)

[2] Workshop 1- School of Computer and Information Technology, Shangqiu Normal University, Henan476000,China; e-mail: ch.x.wang@163.com

trend discovery. Time series are common in all areas of real life, such as finance, meteorology, medical and transportation, etc.. In these time series, there is much valuable information, and the time series association rules are important. Time series sequential association rule mining is a system engineering, which is divided into the time series, the time series compression, the similarity measure of time series pattern, and the acquisition of time series association rules.

For Apriori algorithm needs to scan the database repeatedly, and the Apriori optimization algorithm based on transaction address index table to reduce the transaction is proposed. In this algorithm, an effective method to reduce the transaction is given, which is to reduce the number of candidate sets, and improve the efficiency of Apriori algorithm, but it also needs to scan the database repeatedly.

2. Application analysis of association rules data mining algorithm based on temporal constraint

Data mining can be used to find out the knowledge pattern of decision making, which is found in the given data. It is also called association rules. Association rules are widely used in the field of transactional analysis. In a large number of user's data, there are a lot of association rules, but not all association rules are useful for users.

Its significance lies in the emergence of some of the items in a transaction, can derive other items in the same transaction also appeared (simple (x includes in T) = > (y contains in t said X=>Y. Here in the '= >' called 'Association' operation, X is said for association rules of the prerequisite conditions, y is the result of association rules), as is show by equation (1) [2].

$$\begin{cases} E\left\{v_i(kT_i), v_i^T(jT_i)\right\} = R_i(kT_i)\delta_{kj}, \quad R_i(kT_i) > 0 \\ E\left\{w(jT_i), v_i^T(kT_i)\right\} = 0, \quad E\left\{x(0), v_i^T(kT_i)\right\} = 0 \end{cases} \tag{1}$$

The project set I and transaction database d, I all satisfy the user specified minimum support (Minsupport) of the project, that is greater than or equal to Minsupport I of the nonempty sets, known for frequent item sets or large projects set. Pick out all the frequent itemsets that are not included in the Maximum Itemsets Itemsets Maximum (Frequent) or the largest project set (Large), as is shown by equation (2).

$$L^k = G^k - 4\omega * [G^{k+1}]_{\uparrow 2} \qquad k = 0 \ldots N-1 \tag{2}$$

In order to improve the efficiency of the algorithm, a series of improved algorithms proposed by domestic and foreign experts are mainly from the reduction of the number of scan database and the number of generating candidate itemsets.

Definition 1: temporal item sets: in a project set I of a transaction t is contained, I each time sequence corresponding collection of itemsets called in a transaction t mean set I corresponds to the timing set, with a second tuple < (I), template, said among them, I is a set of projects; the timestamp t is the temporal item set, the value of T is a set I and finally a sequence corresponding to the timestamp.

$$I_k(W(X,P+\Delta P)) = I_k(W(X,P)) + \frac{\partial I}{\partial W}\frac{\partial W}{\partial P} \tag{3}$$

On the multidimensional data model for OLAP operations, rolled up: aggregate data through a dimensional concept hierarchy to rise or by the dimension specification, when used dimensionality reduction of volumes, delete one or more than one dimension by the given data cube, drill down: inverse operation on the volume, by a less detailed data to more detailed data can by concept hierarchy of a dimension along the downward or introducing new dimensionality achieved (more detail is added for the given data) as follows Formula 4.

$$\psi_{a,b}(t) = |a|^{-1/2}\psi(\frac{t-b}{a})b \in R, a \in R, a \neq 0 \tag{4}$$

The least entropy method means value method, boundary value method and median method. (as defined in the data protocol) maintain the integrity of the original data, the data set becomes smaller, does not affect the aggregation (Protocol) method on the results of the analysis data cube; so as to reduce the data mining to deal with the amount of data, improve the mining efficiency (data definition) the data scaling (such as the replacement of large units) to fall into a specific area (such as 0.0 to 1.0), called for standardization. (Common method) the maximum - minimum specification as shown in the following formula (5):

$$\frac{R^{nk}u^{(0)}}{\lambda_1^{nk}} \Rightarrow \sum_{i=1}^{N} \alpha_i \frac{\lambda_i^{nk}}{\lambda_1^{nk}} \tag{5}$$

Definition 2: timing rule support degree and temporal association rules from a to B: T=t in database d with support s, said s is d affairs also contains a, B and meet the percentage of, it is equal to the probability p (AB:] t), s (a - > b:] t) = P (AB:] t) =, where |D|said the number of transactions in a transaction database d; representation of |AB:] t |also contains a, B and meet the transaction number [5].

Classification belongs to the learning, which is a sample study. (data mining is a typical requirement for clustering; the ability to handle different types of attributes; finding arbitrary shape clustering; the ability to determine the domain knowledge of input parameters; the ability of processing the "noise" data; the order of the input record is not sensitive; high dimension; common data type) interval scale variable, proportional scale variable, nominal, ordinal and mixed type. (Dissimilarity matrix) is used to store all of the two two degrees of the dissimilarity between the matrix, for a single NN dimension of the matrix. Its characteristic is d (I, J) =d (J, I), D (I, I) =0, D (J, J) =0.

3. Improved time series Apriori algorithm by frequent itemsets

Apriori algorithm is the use of the association rules. I don't know why, an improved association rule, I think of shopping basket data [6]. This has not been achieved, but it is also understood that it is through the support and confidence to work, but for Apriori, it through the frequent item sets of some of the rules (frequent itemsets of the subset must be frequent itemsets, etc.) to reduce the computational complexity.

So, after scanning these several blocks, we can find the frequency set, and combine the candidate item sets directly. Based on the above reasons, the Apriori algorithm may exist the following problems. (1) There is a large amount of redundancy in the mining rules, as is shown by equation (6). (2) the result of the calculation of an excessive number of computations is slow, the main reason is that the frequent item sets generated too many candidate itemsets, especially the candidate 22 sets are the most serious.

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)(x-\mu)^T}{2\sigma^2}) \tag{6}$$

Definition 3 [7]: in the time series X=$(x_1, x_2, \cdots x_n)$, any two adjacent sub sequences, and, $\left\{\begin{array}{l}|t_2-t_1|<\varepsilon \\ |x_2-x_1|>\delta\end{array}\right.$, $\left\{\begin{array}{l}|t_2-t_1|<\varepsilon \\ |x_2-x_1|>\delta\end{array}\right.$, $\left\{\begin{array}{l}|t_2-t_1|<\varepsilon \\ |x_2-x_1|>\delta\end{array}\right.$ $\left\{\begin{array}{l}|t_2-t_1|<\varepsilon \\ |x_2-x_1|>\delta\end{array}\right.$ to meet, the data has the same increase or decrease trend, but the overall trend is opposite, for a given normal number, there is, as is shown by equation (7).

$$\left\{\begin{array}{l}|t_2-t_1|<\varepsilon \\ |x_2-x_1|>\delta\end{array}\right. \tag{7}$$

Clustering algorithm is a new clustering algorithm which can be used to calculate the association rules. The algorithm can be used in computing the association rules. The algorithm is not only capable of rapid and efficient completion of clustering tasks, but also has a certain ability to deal with multi-dimensional data. The algorithm of particle swarm optimization (PSO) is proposed to optimize the deformation, and the method is suitable for the research.

Definition 4: Given a database of customer transactions, each transaction is composed of a customer's identity$T = (T_1, T_2, \cdots T_k)$, transaction time, and commodity items purchased in the transaction. Transaction data sets, $\bar{x} = Wx =$

$$\begin{bmatrix} x(-s\max,1) \\ \delta x(-s\max,1) \\ \vdots \\ \delta x(-1,\frac{k+N}{2}) \end{bmatrix}_{mN\times 1} \quad \bar{x} = Wx = \begin{bmatrix} x(-s\max,1) \\ \delta x(-s\max,1) \\ \vdots \\ \delta x(-1,\frac{k+N}{2}) \end{bmatrix}_{mN\times 1}$$

respectively, the customer identification, time identification, project set identification.

Definition 5: the use of frequent 1- set with 2-, generating candidate set C2 (Candidate2-itemset). The support count of each candidate item set in C2, the set

L2 of frequent item set 2-, and the combination of 2-, 3-, C3 L2, are generated.

Let I={i1, i2,... Im}, is a set, IK (1 K m) is a sequence of S, S=<s1, s2... Sn>, Sj (1 J N) is a set (S element). Each element is made up of different items. The elements of a sequence can be expressed as (I2, i1,... IK). The sequence contains the number of terms called the sequence of length, the length of the sequence of K is denoted as K – sequence [8]. With alpha =<a1, a2,... An>, beta B2, =<b1,... Bm>, if there is an integer of 1 less than j1<j2<... Jn<m, the A1 BJ1 A2 bj2,... An BJN A1 BJ1, as is shown by equation (8).

$$\bar{x} = Wx = \begin{bmatrix} x(-s\max, 1) \\ \delta x(-s\max, 1) \\ \vdots \\ \delta x(-1, \frac{k+N}{2}) \end{bmatrix}_{mN\times 1} \tag{8}$$

Definition 6: sequence T = <ti1, ti2,... , tim> is another sequence of S=<s1, s2,... And sn> sub sequence of the meet the following conditions: for every J, 1<=j<=m-1, ij<ij+1i and for each a j, 1<=j<=m exist 1<=k<=n. The TIJ SK. That is, sequence S contains T. With the symbol "d" said "is included in" T S, sequence is a sequence of sub sequences can be denoted as T D S. Called T S, the S is a super sequence of T. If a sequence S is not included in any other sequence, the sequence S is the largest.

Definition 7: (frequent item sequence set generation operation) using IS in ISS to select the frequent item sequence and join the (IS, ISS*, ISS) of the frequent item sequence set ISS* (make_fre,):IS* sub {IS}, if IS* support number > minsup_count, IS* may be as frequent item sequence into ISS*.

Since P (X) for all classes as constant, only P (X|Ci) P (Ci) is the largest. If the prior probabilities of the classes are unknown, they are usually assumed to be equal to those of the class. And then the P (Ci|X) is the maximum. P (X|Ci) P (Ci|X), which has a number of attributes, is the P (Ci), which can be very large. To reduce the cost, we can make a simple assumption of class conditional independence. Given the class label of a given sample, if the attribute values are independent of each other, that is, there is no dependency between attributes, as is shown by equation (9).

$$f_u^{(2)} = u^T H u \tag{9}$$

In the Apriori algorithm, the basic idea is to find the maximum set of items: the first step is to simply count the frequency of all items containing one element, and finds out the project set that is greater than or equal to the minimum support degree. From the second step start the loop process until the failure to generate a higher dimension of the frequent item sets.

$$\widehat{q}_b(y_0) = C_1 \sum_{i=1}^{n} k(\left\| \frac{y_0 - \mathrm{X}_i}{h_1} \right\|^2) \delta[(I(\mathrm{X}_i) - b_j], \qquad if \;\; 1 \le j \le m \tag{10}$$

The function of data mining is data mining, which is based on the prediction of the future trend and behavior, and to make the former, based on knowledge.

4. Application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint

A applicable to digital resources access database log association rules mining algorithm, by compressing the affairs and project compression combination, generating candidate item sets at the same time, eliminate transaction database does not support frequent item sets business and affairs of the project, in each provided compressed by a connecting produce candidate item sets and to calculate the degree of support, candidate itemsets by keyword recognition, eliminating the Apriori algorithm in the pruning and string pattern matching step, and it improves the speed of the frequent pattern set. But the algorithm has a narrow application range.

Definition 8: (sequence association rule) for a given item set I={i1, i2,... Im}, T, S sequence, expression form of S T called sequence association rule. With the seed set Lk-1, the candidate sequence set Ck is connected by the 1 - seed set Lk- 1 and Lk-1. Set S1, S2, respectively, the Lk - 1 of the seed set, S2, S1 can be connected to obtain a candidate K sequence of the necessary and sufficient conditions for the S1 to take the first sub sequence of the proceeds obtained from the s2. The rule for generating candidate K sequences is to connect the sequence S1 to the end of the sequence S2. If the last element of the S2 is the last element of the S2, the S1 becomes the last element of the; otherwise, the item becomes the last element of the S1:

$$\Psi_i(m) := diag\left[C_i((m-1)M+1), C_i((m-1)M+2), \cdots, C_i(mM)\right] \tag{11}$$

For a data mining based on data warehouse and data warehouse, the different dimensions of data and the different viewpoints of the data are provided by different dimensions. But it also brings new problems to data mining.

Definition 9: set time series for X=$(x_1, x_2, \cdots x_n)$, arbitrary, there are: $x_i - x_{i-1} \geq 0$, the establishment of the data is with the increase trend; conversely, it is the data with a reduced trend. Inequality is established that is strictly increasing trend data; conversely known is strictly decreasing trend data, where I = 2,3,..., n.

Definition 10: (sequence) I={i1, i2,... , im} is a collection, IK (1=k<=m) is a term, the sequence S S = <s1 =, s2,... And sn> the Sj (1<=j<=n) is a set (also known as sequential element of s), namely I SJ. Each element consists of different, the sequence of elements can be expressed as (I1, I2,... IK), if there is only one sequence in a sequence, the parentheses may be omitted. The sequence contains the number of the number of all the terms called the sequence of length. The sequence of length L is L - sequence.

$$W_\psi f(m,n) = a_0^{-\frac{m}{2}} \int_{-\infty}^{+\infty} f(t)\overline{\psi}(a_0^{-m}t - nb_0)dt \tag{12}$$

There are two types of time series noise data: the first kind is the heterogeneous data which is different from the original data in the time series, such as the time series of computer failure, resulting in different data from the original data in time series, and the second is the data that deviate from the expected value of the sequence. **Definition 11:** set the local extreme points set in the time series X=$(x_1, x_2, \cdots x_n)$ to S=$(S_1, S_2, \cdots S_m)$, and the set of these extreme points are T = $(t_1, t_2, \cdots t_m)$for a given two normal number and if there are two adjacent local extreme points and meet:

$$\begin{cases} |S_k - S_{k-1}| < \varepsilon \\ |t_k - t_{k-1}| < \delta \end{cases} \tag{13}$$

Association rule mining can be used to pre process the temporal constraints, so that we can filter out the user does not care about the time of the data. Filtering database in order to reduce the scan space is the key to reduce the input and output costs, reduce memory requirements and improve the efficiency of mining. A user can use the Iuser mining zone, the above definition of operating T, the realization of the database filtering. Filter (Iuser, D,, D) of the filter operator is described to eliminate the invalid data in D, and eventually generate a new filtered database D.

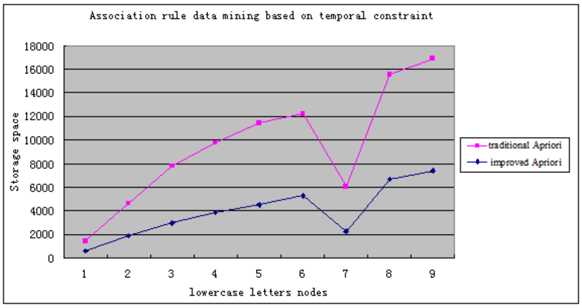

Fig. 1. Comparison of association rule data mining based on temporal constraint based on improved frequent itemsets Apriori algorithm with traditionalApriori

Figure 1 shows that the improved frequent itemsets Apriori algorithm in the storage space is far less than the size of the original transaction database, and relatively stable. The reason is that through the filter out a large number of users do not care about the data (this experiment is about 70%~90%); through the merge to generate a disjoint user time zone (this experiment is about 1~3), the database is scanned for each mining time zone, so that the number of entries into memory is reduced.

5. Conclusions

Thus, it can be used to determine whether the jump degree of time series data is an isolated point. This method has a very strong intuition, which not only takes into account the time series data is a two-dimensional point, but also takes into account the time dimension and the numerical dimension has the very big difference.

The problem of Apriori algorithm is the candidate sequence is easy to generate. The length of the candidate sequences increased by 1, and it was not easy to find a long sequence pattern, because the length of the sequence patterns of the data mining is increased, and the number of candidate sequences is exponential growth with the increase of the length of the sequence patterns in the sequence. The paper presents combined application of improved time series Apriori algorithm by frequent itemsets in association rule data mining based on temporal constraint. Based on this theory, we give an improved algorithm. This algorithm gives the temporal constraint interval, and uses temporal interval algebra to filter and mine the data of the transaction database.

References

[1] J. W. Han, J. Pei, X. F. Yan: *From Sequential Pattern Mining to Structured Pattern Mining:A Pattern- GrowthApproach.* Journal of Computer Science and Technology *19* (2004) 257–279.

[2] N. L. Khobragade, K. C. Deshmukh: *Thermal deformation in a thin circular plate due to a partially distributed heat supply.* Sadhana *30* (2005), No. 4, 555–563.

[3] Y. F. Zhou, Z. M. Wang: *Vibrations of axially moving viscoelastic plate with parabolically varying thickness.* J Sound and Vibration *316* (2008), Nos. 1–5, 198–210.

[4] R. P. Singh, S. K. Jain: *Free asymmetric transverse vibration of parabolically varying thickness polar orthotropic annular plate with flexible edge conditions.* Tamkang Journal of Science and Engineering *7* (2004), No. 1, 41–52.

[5] M. N. Gaikwad, K. C. Deshmukh: *Thermal deflection of an inverse thermoelastic problem in a thin isotropic circular plate.* Applied Mathematical Modelling *29* (2005), No. 9, 797–804.

[6] S. Chakraverty, R. Jindal, V. K. Agarwal: *Flexural vibrations of non-homogeneous elliptic plates.* Indian Journal of Engineering and Materials Sciences *12* (2005) 521–528.

Received November 16, 2016