Journal of Theoretical and Applied Information Technology
20th January 2013. Vol. 47 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN IMPROVED CDAR ALGORITHM BASED ON REDUCING THE SCANNING TRANSACTION DATA

[1] LI-FENG WU, [2] JIAN XIAO

[1] Department of Computer Science and Technology, South Central University for Nationalities, Wuhan 430073, Hubei, China

[2] Hunan University of Commerce Beijin College, Changsha 410219, Hunan, China

ABSTRACT

The research modifies the CDAR algorithm to the itemsets scanning transaction data method, and joins in the concept of deleting the non-frequent itemsets in the transaction data. It judges whether the itemsets can satisfy the minimum support degree. In order to mine the association rules, the paper puts forward an effective algorithm, that is, improved CDAR algorithm. From the experimental estimation, it shows that the improve_CDAR algorithm can effectively improve the execute performance.

Keywords: *Data Mining, Association Rule, Improved CDAR*

1. INTRODUCTION

In the data mining field, how to effectively mine the association rules is one of the mist important research topics. Its purpose is to find out the associations among the product items from many transaction data, and express it in the form of the association rules. As to the enterprise marketing decision, it can offer the useful reference information [1, 2].In the method of mining the association rules, the Apriori algorithm [3] raised by Agrawal and Srikant（1994）is one of the most representative method. It has many features like simple, easily understandable and implemental. In the future research, CDAR algorithm raised by Tsay and Chang-Chien（2004）can effectively improve the execute performance of the Apriori algorithm [4].

The research uses the transaction data as the source of the mining data. Each transaction data includes product items the consumers once purchased. It modifies the CDAR algorithm to the itemsets scanning transaction data method, joins in the concept of deleting the non-frequent itemsets in the transaction data and designs an effectively improving method of mining the association rules, that is, improved_CDAR algorithm.

The frameworks of the paper are as follows: the second chapter introduces the relevant research of mining the association rules; the third chapter explains the process of mining the association rules in the improved_CDAR algorithm, and illustrates it by an instance; the forth chapter estimates the execute performance of the improved_CDAR algorithm by an experiment; at last, the fifth chapter makes a conclusion.

2. RELEVANT RESEARCHES

Agrawal et al.（1993）firstly put forward that it can find out the associations among the product items from many transaction data[5], and expressed it in the form of the association rules. The mining association rules are becoming one of the most important research topics in the data mining field. It explains the relevant definition of the association rules:

It assumes that I is all product itemsets and T is the transaction databases sets of all consumers. There are m sets in total. Each transaction data Tj is sets constituted by several items, $1 \leqslant j \leqslant m$, which is called the itemsets, that is, Tj.I. There are association rules between the itemsets X and Y. It is showed as $X \rightarrow Y$. X is called the antecedent itemsets, and Y is called the consequent itemsets, X 、Y.I and $X \cap Y=$. Whether the association rules $X \rightarrow Y$ is the effective strong rules, it is decided by the two parameters: s and c, that is, support degree and confidence degree. The support degree s is the rate value of instantaneously including the $(X \cup Y)$ in all transaction data sets, namely, s=the transaction data number of including the $(X \cup Y)$/ all transaction data number. The confidence degree c is rate value of instantaneously including Y in the X transaction data sets, namely, c= the transaction data number of including the $(X \cup Y)$/ the transaction data number of including X. The support degree and the confidence degree of the association rules must be greater than or equal to the specified minimum support degree and confidence degree respectively

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

so that association rules can have the significance.

The process of mining the association degrees is composed by two phases: the first phase should find out the itemsets of satisfying the minimum support degree. These itemsets satisfied with the minimum support degree is called frequent itemsets. If an itemsets include K items, it is called the k-itemsets, and it is showed by itemsets. If k-itemsets satisfies the minimum support degree, it is called frequent k-itemsets, and it is showed by frequencies. The second phase counts the association rules formed by the frequencies itemsets in the condition of the minimum confidence degree. If it satisfies the minimum confidence degree, the association rules succeeds. If ABC is frequent 3- itemsets, A、B、C∈I. If the association rules AB→C satisfies the minimum confidence degree, the association rules succeeds.

In the future researches [3, 4, 6, 7, 8, 9, 10], it respectively puts forward different data framework or the data storing method and designs the corresponding mining algorithm for improving the execute performance of mining association rules. Apriori algorithm [3] is the frequencies used method in the mining association rules way and is one of the methods of being used for estimating other algorithm performance. It illustrates the steps of mining association rules in the Apriori algorithm

(1) Find out frequentk-1，k>1. If there is no frequentk-1, k>1, it stops conducting.
(2) The itemsetk is formed by the same frequentk-1 which has two k-2 items in the step (1).
(3) Judge the itemsetk found in the step (2) and check whether the subset of the itemsetk-1 appears in the step (1). If it appears, it should remain the itemsetk; otherwise it is deleted.
(4) Check whether the itemsetk found in the step (3) is satisfied with the minimum support degree. If it is satisfied, it becomes frequentk; otherwise it is deleted.
(5) Count the association rules formed by frequent. If is satisfied with the minimum confidence degree, the association rules succeed.
(6) Skip to the step (1) and then find the frequentk+1 until it will not produce the frequent itemsets.

Document [4] puts forward the CDAR algorithm used to mine the association rules. During the mining process, it includes transaction data whose item numbers are same and these data are clustered into the same group. During the conducting step, it adopts the similar Apriori algorithm as the way of checking whether the itemsetk is satisfied with the minimum support degree, that is, it is unnecessary to scan the transaction data whose item numbers are lower than k cluster. Its purpose is to improve the execute performance of the mining association rules in the Apriori algorithm.

The research modifies the CDAR algorithm to the itemsets scanning transaction data method, joins in the concept of deleting the non-frequent itemsets in the transaction data and improves the execute performance of the mining association rules.

3. MINING ASSOCIATION RULE

If the itemsetk is produced during the process of mining association rules of the CDAR algorithm, the k is greater than 1. Although it avoids the original Apriori algorithm scanning the items whose item numbers are lower than the k transaction data, it must scan the items whose numbers are the same in the transaction data for it can judge whether there belong to the frequent itemsets. Such counting procedure must consume much time in repeatedly scanning the items of the non-frequent itemsets. The chapter uses the transaction data as the source of the mining data, modifies the CDAR algorithm to the itemsets scanning transaction data method and designs an effectively improving method of mining the association rules, that is, improve_CDAR algorithm. The chapter is divided into two parts: the 3.1 chapter raises an algorithm of mining association rules; the 3.2 chapter explains the mining procedure by an example.

3.1 Improve_CDAR Algorithm

During the process of mining association rules in the CDAR algorithm, when it judges whether the itemsets is frequent itemsets, it must scan all items in the transaction data. It can consume much time repeatedly scanning the items of the non-frequent itemsets and lead to the poor performance. In order to avoid the above situations, the paper designs the improve_CDAR algorithm of mining association rules. In the improve_CDAR algorithm, it joins in the concept of deleting the non-frequent itemsets in the transaction data, making the transaction data merely includes the items of the frequent itemsets. It can avoid repeatedly scanning the items of the frequent itemsets and improve the mining execute performance. The process illustrations of the mining association rules in the improve_CDAR algorithm are as follows:

(1) It finds out the frequent1 from the

Journal of Theoretical and Applied Information Technology
20th January 2013. Vol. 47 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

original transaction database D1, and then deletes the transaction data which includes the item number is 1, lastly it forms the transaction database D2.

(2) The itemset2 is formed by two frequentk1 in the step (1). From the database D2, it checks whether itemset2 is satisfied with the minimum support degree. Is it is satisfied, it becomes frequent2; otherwise, it is deleted. In the process of scanning the transaction database D2, it deletes the non-frequent1 items in the transaction data, and then deletes the transaction data which includes the item numbers of the frequent1 is lower than and equal to 2, it forms the transaction database D3 later.

(3) It finds out all frequentk-1， k>2, and then forms the transaction database Dk.

(4) The itemsetk is formed by any two K-2 items which has the same frequent-1 in the step (3).

(5) It judges whether the itemsetk-1 subsets in the itemsetk found in the step (4) appear in the step (3). If it appears, it remains the itemsetk; otherwise, it is deleted.

(6) Scan the transaction database Dk, and check whether the itemsetk found in the step (5) is satisfied with the minimum support degree. If it is satisfied, it becomes frequentk; otherwise, it is deleted. During the process of scanning the transaction database Dk, it deletes all transaction data which includes the item numbers lower than and equal to k, and forms the transaction database Dk+1 later.

(7) Count the association rules formed by frequentk. If it is satisfied with the minimum confidence degree, the association rules succeed.

(8) Skip to the step (3) and then find the frequentk+1 until it will not produce the frequent itemsets.

During each scanning transaction databases, the above algorithm judges whether the itemsetk>1 is in the process of frequent. It remains the spirit of the original CDAR algorithm and avoids scanning the items whose item numbers are lower than the k transaction data. The algorithm can reduce the item numbers of the original transaction data, namely, it just remains the items of the frequent1 itemsets. As to the problems of the itemsetk+1is the count of the frequentk+1, it can effectively scan the transaction data which includes the item numbers. Compared with the CDAR algorithm on the basis of the above counting improvement, the improve_CDAR designed in the paper can effectively find out the association rules.

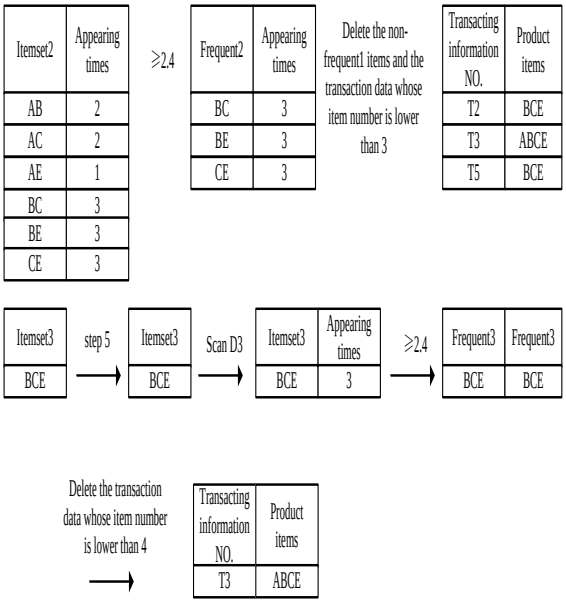
3.2 Experiment Illustration

The paper uses the transaction database D1 as an example. It explains the counting process of using the improve_CDAR algorithm mining association rules. {A, B, C, D, E} are the sets for all product items. {T1, T2, T3, T4, T5, T6} are the sets for 6 consumers' transaction data. Its minimum support degree is 40% (namely, the minimum support number is 2.4) and its minimum confidence is 80%.

Table 1: Transaction Database D1

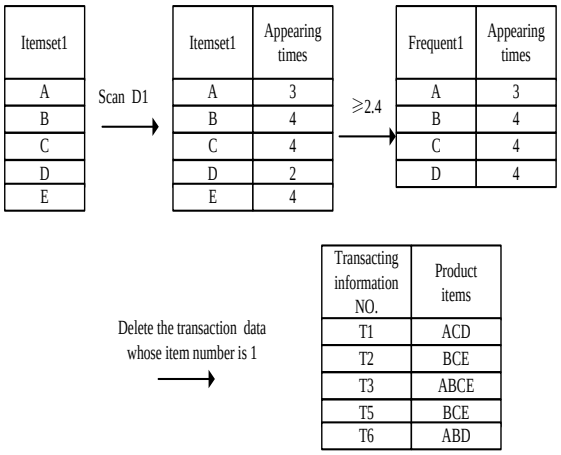
Transaction data NO.	Product items
T1	ACD
T2	BCE
T3	ABCE
T4	E
T5	BCE
T6	ABD

The process of choosing the frequent itemsets is as follows:



Journal of Theoretical and Applied Information Technology
20th January 2013. Vol. 47 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



The paper uses the frequent3-itemsets BCE as an example. It forms the association rules: B→CE, C →BE, E→BC, BC→E, BE→C and CE→B. If the confidence degree is satisfied with the minimum confidence degree, it has the association rules BC→ E, BE→C and CE→B. According to the same counting method, it can find out other association rules formed by the frequent itemsets.

4. SIMULATION EXPERIMENT

The research estimates the execute performance of the improve_CDAR algorithm by an experiment. The illustration of the experimental platform sees in the Table 2. It downloads the data simulation procedure in the IBM Data Mining website (http://www.almaden.ibm.com/) and then the needed transaction data is produced in the estimating experiment.

Table 2: Experimental Platform

CPU	Core 2 dual-core 2.2GHz
RAM	2 GB
Operational system	Microsoft Windows XP Professional SP2
Using Language	C#

The research produces 10 transaction databases which include 10k transaction databases respectively, and then accumulates the front 10 transaction databases in sequence, whose numbers are 10k, 20k, 30k, …, 100k transaction databases respectively, and they are showed by the serial number D1, D2, D3, D4, D5, D6, D7, D8, D9, D10, seen in the Table 3. They are used for estimating the execute performance of the transaction databases. The meaning of the mainly parameter values in the transaction databases are: n represents the number of the items; ntran is the number of the transaction data; np is the number of the combination pattern; tl is the average item number of the transaction data; pl is the average length of the frequent itemsets; the rest parameters is showed by the default values. In the process of mining the computer, it sets the minimum confidence degree as 80%, and then estimates the execute performance of improve_CDAR algorithm and CDAR algorithm respectively.

Table 3: Transaction Databases And Its Parameters

Parameter / Database	n	ntran	np	tl	*pl*
D1	1000	10k	10000	20	*10*
D2	1000	20k	10000	20	*10*
D3	1000	30k	10000	20	*10*
D4	1000	40k	10000	20	*10*
D5	1000	50k	10000	20	*10*
D6	1000	60k	10000	20	*10*
D7	1000	70k	10000	20	*10*
D8	1000	80k	10000	20	*10*
D9	1000	90k	10000	20	*10*
D10	*1000*	*100k*	*10000*	*20*	*10*

The figure 1 uses the transaction database D5 as the source of the mining data. It estimates the execute time of the improve_CDAR algorithm and CDAR algorithm in the condition of the different minimum support degree. The fixed minimum support degree in the figure 2 is 0.012. It uses the transaction database D1, D2, D3, D4, D5, D6, D7, D8, D9, D10 as the source of the mining data, and then estimates the execute time of the improve_CDAR algorithm and CDAR algorithm.

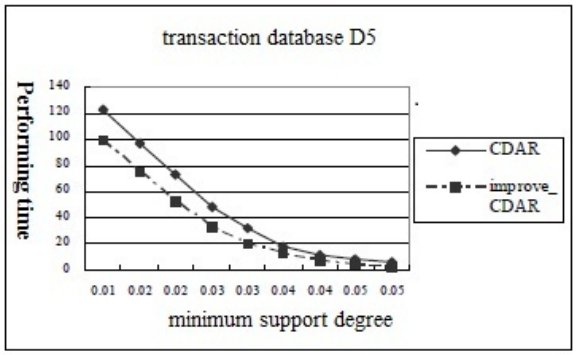


Figure 1: Performing Time Of The Different Minimum Support Degree

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

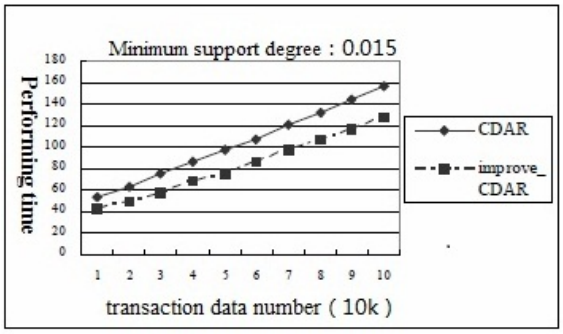


Figure 2: Performing Time Of The Different Transaction Data Number

It shows that the execute performance of the improve_ CDAR algorithm designed in the paper can effectively improve the mining association rules in the CDAR algorithm from the above experimental estimation.

5. CONCLUSION

Improving the execute performance of mining association rules in the mining data field is one of the most important researches. Judging whether the itemsets is frequent itemsets in the mining process must consume considerable counting time in scanning the transaction data. Therefore, if it can reduce the number of each scanning transaction data and the number of the items in the transaction data, it is helpful to improve the mining performance. The research uses the transaction data as the source of the mining data, modifies the CDAR algorithm to the itemsets scanning transaction data method, joins in the concept of deleting the non-frequent itemsets in the transaction data and designs an the mining association rules of the improve_CDAR algorithm. Form the experimental estimation, the execute performance of the improve_CDAR algorithm is better than CDAR algorithm.

REFRENCES:

[1] J. Abbott, M. Stone, and F. Buttle, "Customer Relationship Management in Practice – A Qualitative Study", *Journal of Database Management*, Vol. 9, No. 1, 2001, pp. 24–34.

[2] Jigna J. Jadav, Mahesh Panchal . "Association Rule Mining Method On OLAP Cube", *International Journal of Engineering Research and Applications (IJERA)*, Vol. 2, No. 2, 2012, pp.1147-1151.

[3] R. Agrawal and R.Srikant, "Fast Algorithms for Mining Association Rules in Large Database," Proceedings of the 20th International Conference on Very Large Data Bases, September 12-15, 1994, pp. 487-499.

[4] Y. J. Tsay and Y. W Chang-Chien, "An Efficient Cluster and Decomposition Algorithm for Mining Association Rules," *Information Sciences*, Vol. 160, 2004, pp. 161-171.

[5] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Very Large Database," Proceedings of the ACM SIGMOD Conference on Management of Data, May 26-28, 1993, pp. 207-216.

[6] Ramesh C. Agarwal, Charu C. Aggarwal, V.V.V, "A Tree Projection Algorithm for Generation of Frequent Itemsets," *Journal of Parallel and Distributed Computing*, Vol. 63, No. 3, 2000, pp. 350-371.

[7] F. Coenen, P.Leng and Ahmed, S. Ahmed, "Data Structure for Association Rule Mining: T-trees and P-trees," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 6,2004, pp. 774-778.

[8] J. Han, J. Pei, Y. Yin and R. Mao, "Mining Frequent Patterns without Candidate Generation: a Frequent-Pattern Tree Approach," *Data Mining and Knowledge Discovery*, Vol. 8, No. 1, 2004, pp. 53-87.

[9] J. D. Holt, and S. M. Chung, "Mining Association Rules Using Inverted Hashing and Pruning," *Information Processing Letters*, Vol. 83, No. 4, 2002, pp. 211-220.

[10] Z. C. Li, P. L. He and M. Lei, "A High Efficient AprioriTid Algorithm for Mining Association Rule," Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, August 18-21, 2005, pp. 1812-1815.