**TELKOMNIKA**, Vol.11, No.11, November 2013, pp. 6521~6526
e-ISSN: 2087-278X



# An Improved Apriori Algorithm for Association Rules

**Xingli Liu*, Huali Liu**
Department of Computer and Information Engineering, Heilongjiang Institute of Science and Technology, Harbin 150027, Heilongjiang, China
*Corresponding author, e-mail: liuxingli_usth@139.com

***Abstract***

*According Apriori algorithm characteristic achieve its improvement and apply it to the knowledge correlation of the curriculum in sumulation experiment. Firstly, it is mainly by simplifying the binary storage method to change data in the database, and then to get the largest frequent itemsets. The experiment results showed that the improved algorithm obviously improve the efficiency; secondly, establish a new database to simulate applied experiment, consisted of student achievement of various knowledge points in the computer programming course, and then using this optimized algorithm to found the course knowledge frequent itemsets in a database, which is closely interrelated knowledge points mainly by setting up different minimum support value to get various frequent itemsets. According to these frequent itemsets of the course it can be applied to reestablish a new course knowledge system to further improve the teaching quality, this method can also be achieved the knowledge system reform of other course or course group.*

***Keywords:*** *apriori algorithm, association rules, curriculum knowledge correlation*



## 1. Introduction

Data mining is a subject of the application in the field of endeavors which to provide policy support for the fields based on information analysis. Data mining extract the unknown, implied and useful information and knowledge from a large number of structural and the structure of data. As one of the most popular method of data mining which signify local patterns, association rules provide a very simple and valuable description. Apriori algorithm is the most classic one to found frequent item set [1].

Nowadays, there are a lot of impoved Apriori algorithm.For example,according to the improved Apriori algorithm, which scans the database only once by using arrays to store data and the new algorithm sorts the frequent itemsets from small to large [2]. For instance, Coenen and Leng proposed Apriori-TFP (Total-from-Partial) based on Apriori improvment, which used to P (partial sup-port tree) and T (total support tree) to decrease computer time and storage space [3], and as literature 4 proposed to improve the the way of candidate connection [4], and as others put forward the the improvement of the algorithm [5, 6]. After that improved Apriori algorithm by appied widly in the aspects. For example, Literature 11 the improved algorithm can quickly find the corelation of the stock plate, it has a certain guiding role in the stock market analysis and investment decisions [7]. As literature 12 popose a model of intrusion detection system based on class-association rule to improve the detection rate of intrusion detection system, experiments show that the proposed method could detect intrusions efficiently in the network [8], and there is a aviation safety reports analysis application and web recommender system in another literature [9, 10].

Of course, improved Apriori althorigthm is also put into use in teaching, as literature propose the measure quality of teaching [11]. But there is less in the curriculum. In the teaching process, there should have the small tests to control the situation of the students. Therefore, there is large number of students in the usual results of the feedback at every semester or the end of the year. However, some important potential value has not been utilized. For example, there are close touch with comparison in the final stages of review which can be connected to a review class in a more easy to learn knowledge framework to review. This can also guide the next instruction, it is likely to cause other links between the teaching processes, and we should pay attention to the strengthening of knowledge leading. Comprehensive, we hope to find a course in connection with the close to where it is more and more conducive of the structures to

*Received March 15, 2013; Revised June 18, 2013; Accepted July 19, 2013*

construct the course of the intellectual system. Therefore, we can dig up by the rules of the largest number of project assembly to complete. But the traditional rules which are iterated to produce the maximum number of project assembly are not entirely suitable. So in accordance with computer programming charateristic this paper achieve knowledge system establish based on analysis of the classical Apriori algorithm in mining association rule. The improved algorithm adopts matrix to express database, just scans database once, cancels a great number of linking operations in finding frequent item sets dimension by dimension but finds out the highest dimension frequent item sets directly from high dimension itemsets, the algorithm is improved efficiently.

## 2. Research Theory

In 1993, Agrawal and several people are the first carried up with to excavate the database transaction set of rules and associated issues to find between different goods of the rules, then they put forward Apriori algorithm [12]. These basic concepts are below.

Definition 1 is item: a field in a transaction database, it usually uses the lowercase $i_m$ as the mark. For example, basic knowledge is the Item in the C++ Programming.

Definition 2 is transaction: Correspond to a record in the course results database, it usually uses the lowercase ti as the mark, ti= {il, i2, …, ip}. Every transaction has a unique identifier that calls TID. There are a transaction such as {basic knowledge of C++, basic knowledge of object-oriented, destructor and constructor, and so on }.

Definition 3 is Itemset: The database for all the items collection mark the upper I, the any subset of x is called the itemset of D, it means that equivalently the set of i1, i2, …, im.

Definition 4 is dimensionality of itemset: the number of items which Itemset contains, it will be called K-Itemset If the number is K.

Definition 5 is association Rules: if both X and Y are itemsets, and X∩Y=φ, then there is association rules X=>Y, and its meaning that X has also led to y, X is premise and Y is conclusion.

Definition 6 is support: there is X=>Y among X belong to I and Y also belong to I, at the same time X∩Y=φ, the support of Rule is support (X=>Y). It also means on the deal set at once the percentage of X and Y, such as formula (1).

$$support(X \Rightarrow Y) = \frac{|\{ti \mid X \cup Y \subseteq ti, ti \in D\}|}{|D|} = p(X \cup Y) \quad (1)$$

Definition 7 is confidence: there is X=>Y among X belong to I and Y also belong to I, at the same time X∩Y=φ, the confidence (X=>Y) of Rule is the proportion of transactions number of include of X,Y and X. such as formula (2).

$$confidence(X \Rightarrow Y) = \frac{|\{ti \mid X \cup Y \subseteq ti, ti \in D\}|}{|\{tj \mid X \subseteq tj, tj \in D\}|} = \frac{support(X \cup Y)}{support(X)} = p(Y \mid X) \quad (2)$$

Definition 8 is the minimum support and the minimum confidence of association rules, The former, namely $sup_{min}$, is measure the lowest importance of rules requirement and the latter, namely $conf_{min,}$ is he minimum reliability of rules requirement.

Definition 9 is strong association rules, if rules is X=>Y, and meet support (X=>Y)>=$sup_{min}$ and confidence (X=>Y)>= $conf_{min}$ , then this rules is strong association about X=>Y,otherwise is weak association.

According to the classical Apriori algorithm, There are two most important quality for associated rules: The first one is that the set of frequent Itemsets is frequent Itemsets, the other one is that the superset of unfrequent Itemsets is unfrequent Itemsets. In 1994, Agrawal is the first carried up with the famous Apriori algorithms. The process as follows.

Input: transactional databases D and the minimum support is $sup_{min}$.

Output: all the frequent itemsets $L_i$ in the D.

(1) L1=find_frequent_1_itemsets(D);

(2) for (k=2;$L_{k-1}$≠ Φ;k++){

(3) $C_k$=apriori_gen ($L_{k-1}$, $sup_{min}$);

(4) for each t∈D {

(5) $C_t$=subset ($C_k$, t);
(6) for each c∈$C_t$ c.count++;
(7) }
(8) Lk={c∈Ck|c.count> supmin };
(9) return L= $\cup_k L_k$;
The 3th step above the algorithms of apriori_gen ($L_{k-1}$, $sup_{min}$) described below.
Input: last scan-round result $L_{k-1}$ and the minimum support is $sup_{min}$.
Output: candidate frequent itemset $C_k$.
(3.1) for each l1∈$L_{k-1}$
(3.2) for each l2∈$L_{k-1}$
(3.3) if((1l[1]=l2[1])∧…∧(1l[k-2]=l2[k-2]) ∧(1l[k-1]<l2[k-1]) ) {
(3.4) c=$l_1 \oplus l_2$;
(3.5) if has_infrequent_subset(c, $L_{k-1}$)
(3.6) delete c;
(3.7) else $C_k$=$C_k \cup$ {c};
(3.8) }
(3.9) return $C_k$;
The 3.5th step above the algorithms has_infrequent_subset(c,$L_{k-1}$) described below.
Input:this scan generate the each of $C_k$ subset c,last scan generate $L_{k-1}$.
Output: whether or not c is deleted from $C_k$.
(3.5.1) for each (k-1)-subset s of c
(3.5.2) if s∉$L_{k-1}$ return TRUE;
(3.5.3) else return FALSE;

## 3. Improvment Results and Analysis

The classical Apriori algorithm used k of frequently items to generate K+1 candidate in connections through cutting to have frequent and a set of items, until not frequent items. Apriori algorithm is generated every length of the assembly that will scan a database, and to be a candidate, so the number of scans the database are decided by the most frequent set of items. So the paper improves a new algorithm against the problem.

### 3.1. Improvment Method

There is a great of research about Apriori algorithm in improvement. To sum up, if the number of K frequent itemset, it is $L_K$ <=K,this set is the largest frequent itemset which less or equal to K, it is support the maximum items number all the transaction, so in the improvment algorithm the largest frequent itemset k, according to above what the conclusions to determine, is used, through only pay attention to the affairs of item number which greater or equal to k to find the largest frequent $L_k$. If not find frequent itemset K then k-1, k is the hypothetical item number of the largest frequent itemset, and then repeated according to the above method. There is an analog transactional database as Table 1, each of the transaction T is a student record, from I1 to I7 represent different knowledge, $sup_{min}$ is 4.

Table 1. The Original DataBase

| Transaction | Item | Flag1 |
|---|---|---|
| T1 | I1,I7 | 2 |
| T2 | I2,I3,I4 | 3 |
| T3 | I4,I5,I7 | 3 |
| T4 | I2,I4,I5 | 3 |
| T5 | I5,I6,I7 | 3 |
| | …… | |
| T15 | I4,I5,I7 | 3 |
| T16 | I3,I4 | 2 |
| T17 | I1,I3,I5,I7 | 4 |
| T18 | I2,I4,I5,I6 | 4 |
| T19 | I5 | 1 |
| T20 | I2,I4,I5,I6 | 4 |

When Scanning the database record C[n] there is C[1] is 1, C[2] is 5, C[3] is 8, C[4] is 4, C[5] is 2, then set each of the transaction Flag1 and change data storage in order to item as

keyword, whether or not transaction include of this item by 0 or Flagi. If every transaction the maximum length is n, then use n binary to represent its transaction set, this n system storage by item as keyword shown in the following Table 2.

Table 2. The n System Storage by Item as Keyword

| Item | Transaction | Flag2 |
|---|---|---|
| I1 | 2 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 4 0 0 0 | 3 |
| I2 | 0 3 0 3 0 2 0 4 0 3 0 5 5 3 0 0 0 4 0 4 | 10 |
| I3 | 0 3 0 0 0 0 3 0 0 0 0 5 0 3 0 2 4 0 0 0 | 6 |
| I4 | 0 3 3 3 0 2 0 4 0 3 2 5 5 3 3 2 0 4 0 4 | 14 |
| I5 | 0 0 3 3 3 0 3 4 2 3 0 5 5 0 3 0 4 4 1 4 | 14 |
| I6 | 0 0 0 0 3 0 0 4 0 0 0 0 5 0 0 0 0 4 0 4 | 5 |
| I7 | 2 0 3 0 3 0 3 0 2 0 2 0 5 0 3 0 4 0 0 0 | 9 |

If set C[4]+C[5]> $sup_{min}$, then presuppose the maximum frequent itemset number is 4. Beacause $sup_{min}$ is 4, the Flag 2 of I1 is 3, it meaning there is three transactions in the I1, so I can't be frequent itemset, of course, it isn't the largest one. Now it should only to judge I2, I4, I5, I6, I7 after delete I1 and its data record. Presuppose that the maximum frequent itemset number k is 4, then if the transaction item is equal or greater k, then change transaction value to 1, or 0. This temporary binary database shown in the following Table 3.

Table 3. The Temporary Binary Database

| Item | Transaction | Flag2 |
|---|---|---|
| I2 | 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 1 0 1 | 5 |
| I3 | 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 | 2 |
| I4 | 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 1 0 1 | 5 |
| I5 | 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 1 1 0 1 | 6 |
| I6 | 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 1 | 4 |
| I7 | 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 | 2 |

There are olny the transaction of I2, I4, I5, I6 which supmin is satisfactory, then solve the mixed set of I2, I4, I5, I6. Now that the data as binary to store, so it will be finished with ∩operation of binary. If number of elements is equal or greater to minimum support, then deduce the set of I2, I4, I5, I6, it is the largest frequent itemset, and then In the set of 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 1 there are involving four 1, it's conform to the minimum support. Usually, with the assumption that the maximum number of items frequent itemsets k equals the number of items. Typically, in the solution process, such as seeking I2, I4 intersection set. Their transaction$\cap_{2,4}$, if the number "1" is less than 4, that does not meet the minimum support required, then the number of items for consideration when less than k, there is no maximum frequent itemsets, then do not need to continue to seek common ground. If you do not find 4 frequent itemsets, then it will assume the item number of frequent item sets the maximum number of items k minus 1, in accordance with the above method to obtain the same. According to the nature of Apriori algorithm, a subset of frequent itemset are frequent itemset, so the maximum frequent itemset: I2, I4, I5, I6 to produce some two frequent itemset: I2, I4 and I2, I5 and I2, I6 and I4, I5 and I4, I6 and I5, I6, there are there frequent itemset: I2, I4, I5 and so on.

### 3.2. Improved Algorithm Efficient Analysis

According to the method of the above improvement Arpriori algorithm, the experimental simulation data from C++ examination in the computer programming. Improve the efficiency of the algorithm compared before and after. As shown in the Figure 1, this algorithm is the number of items in the mining maximum frequent set, no access to all databases, but only concerned about the affairs of the number of items that is greater than the number of items which is equal to the assumed maximum frequent item sets the number of items. Meanwhile, not the transaction as the keyword, but the project as a keyword on data storage, and simplified step by step, finally stored in binary form, which can not only save storage space, but also through the use of binary "and" obtain the maximum frequent item sets operation simplifies the algorithm, thus enhancing the efficiency of the algorithm.

### 3.3. Improved Algorithm Application Found

In the C++ programming examination in the computer by the VC. There are 10 knowledges, respectively: c + + preliminary knowledge, object-oriented basic knowledge of the constructor and destructor, the class static members, class friend, class templates, operator overloading, inheritance and derived more state and virtual functions, input and output streams. Then each student has a total score, in accordance with the Arabic numerals 1-11 sequential number. Each knowledge point first determine the passing score, if the student has mastered a certain knowledge point, then the knowledge of the corresponding point in the student affairs records may correspond to the database as follows, as Table 4.

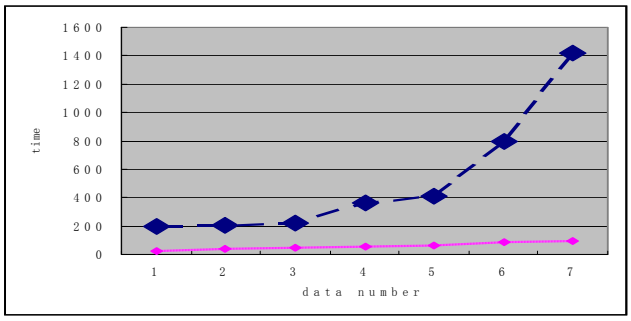


Figure 1. The Comparison of between before and after

Table 4. Simulation Database of Knowledge Scores

| Transaction | Item | Flag1 |
|---|---|---|
| T1 | I1,I2,I3,I4,I5,I7,I8,I9,I11 | 9 |
| T2 | I1,I2,I5,I7,I8,I9,I11 | 7 |
| T4 | I2,I3,I5,I7,I8 | 5 |
| T5 | I2,I3,I4,I5,I7,I8,I9,I10,I11 | 9 |
| T6 | I1,I2,I3,I4,I5,I7,I8,I9,I10,I11 | 10 |
| T7 | I1,I2,I3,I5,I7,I9 | 6 |
| T8 | I1,I2,I5,I7,I8,I9,I11 | 7 |
| | …… …… | |
| | …… …… | |
| T53 | I1,I2,I3,I4,I5,I7,I8,I9,I10,I11 | 10 |
| T54 | I1,I2,I3,I5,I6,I7,I8,I9,I10,I11 | 10 |
| | I1,I2,I3,I4,I5,I6,I7,I8,I9,I10,I1 | |
| T55 | 1 | 11 |
| T56 | I2,I3,I5,I6,I7,I8,I9,I11 | 8 |
| T57 | I1,I2,I3,I5,I6,I7,I8,I9,I10,I11 | 10 |
| T58 | I1,I2,I5,I7 | 4 |
| T59 | I1,I2,I3,I5,I6,I7,I8,I9,I10,I11 | 10 |
| T60 | I2,I7 | 2 |
| T61 | I2,I3,I4,I5,I7,I8,I9,I11 | 8 |
| T62 | I2,I4,I5,I7,I8,I9,I11 | 7 |

Table 5. Simulation Result

| Min_support | Maximum frequent itemsets | Min_support | Maximum frequent itemsets |
|---|---|---|---|
| 37 | I2、I3、I5 | 43 | I2、I3、I9 |
| | I2、I3、I7 | | I5、I7、I8 |
| | I2、I3、I8 | | I5、I7、I9 |
| | I2、I3、I11 | | I7、I8、I9 |
| | I3、I5、I7 | | I7、I8、I11 |
| | I3、I5、I8 | | I8、I9、I11 |
| | I3、I5、I9 | | I2、I5 |
| | I3、I5、I11 | | I2、I7 |
| | I5、I7、I9 | | I2、I8 |
| | I5、I7、I11 | | I2、I11 |
| | I7、I8、I9 | | I3、I8 |
| | I7、I8、I11 | 50 | I2、I7 |
| | I8、I9、I11 | | I2、I9 |

When Min_support value is not the same time, you can frequent itemsets results obtained knowledge of C++ courses point set, as Table 5 shown. When the minimum support in mind Min_support=43(70%*62), we can draw frequent item sets with 8; When the minimum support in mind Min_support=50(80%*62), we can draw frequent item sets have two. According to Apriori definition 6, we learn that each maximal frequent item sets are closely related to the knowledge of point set. Meanwhile, with the Min_support value increases gradually, reached the maximum frequent are concentrated close degree of knowledge points higher. Therefore, in the teaching process, will be closely related to the point of contact with the study of knowledge will improve the efficiency of learning.

## 4. Conclusion

Classic Apriori algorithm for association rule mining algorithm has been widely used but there have been many deficiencies of the algorithm for the improved algorithm. This paper primarily aims to find the association of knowledge points to close-knit comprehensive review at the end of the semester curriculum review, and build a better review system and guide future teaching. Apriori algorithm in the study of theory and extensive literature, based on the number of items for mining frequent itemsets put forward an improved algorithm. Meanwhile, the data storage also changes accordingly. Finally, the improved algorithm experimental results show that the improved algorithm is indeed a higher efficiency, improved performance of the algorithm. In addition, the algorithm can also be applied to teaching other aspects of the monitoring and early warning to mining, such as, to provide a scientific algorithm support for the employment of college graduates on early warning; undoubtedly, the algorithm improvements and application in the curriculum is important value.

## Acknowledgements

This work was supported by 2013 Heilongjiang Province Education Department of humanities and social science research projects. Theory and practice of employment of college graduates monitoring and early warning mechanism (Issue Number：1253xs106). We would like to acknowledge issue's sponsors for its financial support, and as well as thank the reviewers for their valuable feedback.

## References

[1] Chen MS, Han JW, Yu PS. Data M in ing: An Overview from a Database Perspective. *IEEE Transactionson Knowledge and Data Engineering*. 1996; 8(6): 866-883.
[2] Bi Xujing, Xu Weixiang. The Research of Improved Apriori Algorithm. LISS 2012.2013; 1007-1012.
[3] Coenen F, Leng P, Ahmed S. Data structures for association rulemining: t-trees and p-trees. *IEEE Transactions on Data andKnowledge Engineering*. 2004; 16(6): 774-778.
[4] Liu H uating Guo R enx iang Jiang H ao. Research and Improvement of Apriori Algorithm for Mining Association Rules. *Computer Applications and Sofwtare*. 2009; 26(1): 146-149
[5] H Toivonen. Sampling Large Databases for Association Rules. *In The VLDB Journal*. 1996; 134–145.
[6] JS Park, MS Chen, PS Yu. *An Effective Hash Based Algorithm for Mining Association Rules*. Proceedings of the 1995 ACMSIGMOD International Conference on Management of Data. San Jose, California. 1995; 22–25.
[7] Zhang Jianlin, Zhou Chaoliang. Application of Association Rules in Stock Plate Cointegration Analysis. *Computer Engineering and Applications*. 2013; 492(2): 242-245.
[8] CAI Wei-xian, TENG Shao-hua. Application of improved Apriori-TFP algorithm in intrusion detection. *Computer Engineering and Design*. 2011; 32(11): 3594-3598.
[9] Fang Xia, LI Juan-juan, YAN Guan-nan. Application of association rules mining in aviation safety reports analysis. *Computer Engineering and Design.* 2011; 32(1): 218-221.
[10] Xie Nannan, Hu Liang, Nurbo, Fan Li, Yin Xiao tian. Web Recommender System Banse on Web Log Mining. *Jouranl of Jinlin University (Science Edition)*. 2013; 51(2): 267-272.
[11] Song Xiaomei. Data Mining Technology in the Application of the University Teachers' Teaching Quality Assessment. *ChaFeng Collegde Journals.* 2013; 29(2): 181-184
[12] Agrwal R, Srikant R. *Fast Algorithms for Mining Association Rules in Large Ddatabases*. Proceedings of the Twentieth International Conference on Very Large Databases. Santiago, Chile. 1994; 487-499.