The New Algorithms of Weighted Association Rules Based on Apriori and FP-Growth Methods

Ting Liu
College of Information Science and Technology, Zhengzhou Normal University,
Henan Zhengzhou, 450044, China
E-mail: eduliuting@163.com

Abstract

In order to improve the frequent itemsets generated layer-wise efficiency, the paper uses the Apriori property to reduce the search space. FP-grow algorithm for mining frequent pattern steps mainly is divided into two steps: FP-tree and FP-tree to construct a recursive mining. Algorithm FP-Growth is to avoid the high cost of candidate itemsets generation, fewer, more efficient scanning. The paper puts forward the new algorithms of weighted association rules based on Apriori and FP-Growth methods. In the same support, this method is the most effective and stable maximum frequent itemsets mining capacity and minimum execution time. Through theoretical analysis and experimental simulation of the performance of the algorithm is discussed, it is proved that the algorithm is feasible and effective.

Keywords: FP-grow, apriori, weighted association rule

1. Introduction

Data mining association rule mining is an important research topic in the field. Association rules can generally be divided into Boolean association rule and quantitative association rules. Agrawal in 1993 made Boolean association rules is proposed, after classics Apriori and Apriori TID algorithm. Multi valued attributes are divided into categories of attributes and attribute, many algorithms in solving multi valued attributes mining association rules, are the continuous numerical discrimination, get the corresponding fuzzy description, and its processing method is similar to the Boolean association rules mining.

As in the weighted association rules based on simultaneous mining positive and negative association rules, will produce some contradictions and meaningless rules, therefore, in the traditional support, confidence framework, introduced in third parameters to remove the redundant rules. Based on the correlation between positive and negative association rules mining algorithm based on interest and right: the positive and negative association rules mining algorithm based on weighted chi-square test; the positive and negative association rules mining algorithm [1]. When the database is very unevenly distributed, the above for the mining of association rules is not effective. Because of the low frequency items whose support is often low, and therefore rarely be exhumed. Aiming at this problem, the support degree model for different transactions with different minimum support threshold, it is more efficient user interest rules.

Weighted association rules with negative support and confidence of the calculated by introducing negative set, support degree and confidence degree calculation is facing a severe test of its search space exponential growth. Because data mining is the object of mass data, and association rules are in focus for the project of it.

Class with Apriori algorithm based on FP-Growth algorithm; a class based on a class of algorithms. Previous research at home and abroad have done a lot of work, their typical algorithms are proposed in MINWAL (O) algorithm and MINWAL (W) algorithm, the proposed algorithm New-Apriori 2003, in addition to WAR algorithm and the MWQAR algorithm. The paper puts forward the new algorithms of weighted association rules based on Apriori and FP-Growth methods.

2. The Analysis of Weighted Association Rules Technology

With the rapid development of computer technology, the society has entered the information age, especially the database and data collection technology development, various industries store a lot of data, from the original file data to the computer storage. Because the database expansion capacity, large amounts of data in database, much important information behind, and it is the information is extracted from the database, will create a lot of potential value. In the face of the increasing expansion of sea of data, the traditional data analysis methods have been unable to meet the needs of the people, as is shown by Figure1.
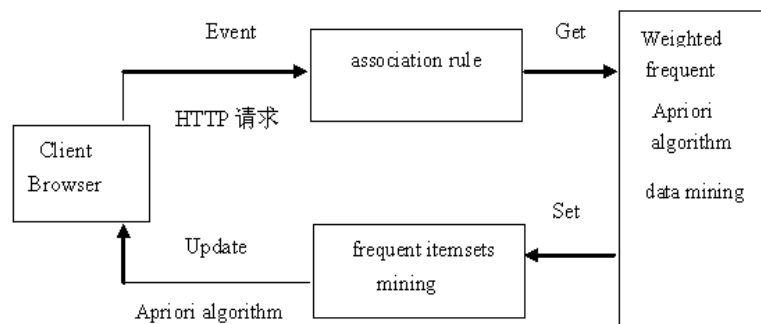

Figure 1. Weighted Association Rules Technology Diagram

Since the association rules in data mining, first by Agrawal was put forward, people on the association rules in data mining technology research has been conducted, in theory it on a lot of very fruitful analysis and research, practice also made many effective algorithms for mining association rules from theory to application foundation. Here, based on data mining association rules algorithm introduction. Association rule mining algorithm classic algorithm is Apriori algorithm. The algorithm is proposed by Agrawal et al in 1994, the main work is on association rule frequent item set mining, especially for Boolean association rules frequent itemsets mining.

It can be attributed to the following two kind of thought: one is the attribute discrimination, it is weighted, the problem is transformed into a weighted Boolean Association rules. The other method is the attribute of domain is divided into overlapping interval, then located near the boundary element makes it possible to simultaneously in the two intervals. As a result of these elements simultaneously to the two zones between contributions, there may be too much emphasis on the role of these elements.

The eight parameter perspective transformation model although it can more accurately describe the camera's movement, but its computational complexity greater degree. Considering the accuracy with real-time requirements, it is the six-parameter affine model of camera motion caused by the interface scene change modeling [2]. When the scene relative depth change is not significant, the six-parameter affine model can well describe the rotation of the camera, panning and translational motion.

At present in the association rules algorithm for frequent itemsets discovery, has produced various effective method. Generally speaking, these algorithms are to follow two steps: one, to establish the frequent item sets a candidate set; two, in the candidate set out actually includes frequent item set of all subsets. In association rule method, Apriori method is most widely used. The Apriori method is used as a layer by layer search iterative method, and it is using the previous iteration of the frequent item sets as after one is iteration of the candidate item sets, and then gets the final result prune.

Wherein, each transaction is a set of T is each transaction having an identifier TID. Let A be a set, T contains A when the association rules are shaped as A→ B implication, which, A I, and only when the A T B, I, A∩ B=. Rule A→ B in transaction set D support s D transaction contains A→ B percentage, let menus be the minimum support degree, if s ≥minsup, called for frequent itemsets. Rule A→ B in transaction set D confidence level in C is contained in the D A transaction contains both the percentage of B, namely Support (A→ B) =P (A ∪B), Confidence (A→ B) = P (A|B), as is shown by Equation (1).

$$f(u_0,v_0) = f(u_0,v) + \beta\left[f(u_0,v+1) - f(u_0,v)\right]$$
$$= (1-\alpha)(1-\beta)f(u,v) + \alpha(1-\beta)f(u+1,v) +$$
$$(1-\alpha)\beta f(u,v+1) + \alpha\beta f(u+1,v+1)$$

(1)

Weighted frequent itemsets downward closure character in traditional association rules in transaction, a lot of, thus creating enormous subset, using frequent set the downward closure character, to some rules pruning, for example, in the algorithm Apriori, if {AB} and {BC} are not frequent, then {ABC} and {BCD} are not frequent.

Improved weighted association rules algorithm and Apriori algorithm the basic steps like: first find all weighted support degree is not smaller than a user-specified minimum support degree weighted frequent itemsets, and the frequent itemsets satisfying the minimum weighted confidence of all the rules.

A set of {AB} affairs weight calculation for: For (k=2; L k-1 = do begin; k++) Ck =Apriori-gen (L k-1); For each transactions t∈D For each transactions t D {Ct∈subset (Ck, t); For each candidates C ∈Ct{ c.num++; c.weight= Weight (c); / / calculation of weighted support } Lk ={c ∈Ck | c. weight ×c.num/|D|≥ minwsup} } return L= ∪K Lk; ITW (A, B) =0.6 ×0.9/ (0.6+0.9) =0.36. Weighted support database contains the project affairs centralization value summary : WS (X) =N X ∏([i[w] ∈X]) Ti [ik [w]] n Σ([i[w]∈ X Ti [ik [w]]])K =1 K =1 |X |X |.

Because the data the body itself is not too much significance, it is only a description of what happened, and not for decision-making provide reliable basis; through data analysis to find out the relationship between the given data, some meaning and relevance, it formed the so-called information. Although given data in some meaningful things, but it often and people need to complete the task without direct contact, also cannot serve as decision-making basis; only the information for further processing, to undertake more thorough analysis, in order to obtain more useful information, namely knowledge, as is shown by Equation (2) [3].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}[R(x_i,y_j) - F(x_i,y_j)]^2}{M \times N}}$$

(2)

However, the amount of data to the explosive growth of now makes it difficult for the user to like before relying on experience, large amounts of calculation and human command to find out about the artificial data more comprehensive knowledge, the knowledge hidden in the data. Many can be found and used, the waste of resource caused by data. Many decision makers from the database mining these patterns are increasingly interested in, association rules mining can provide effective decision support, was promoted on certain level association rules mining technology development.

The problem of mining association rules can be formalized as follows: I}, is the collection of all items. Is all affairs D M set (database), each transaction T is the number of items collection, T included in I, each transaction can use T I to identify a unique identifier. A D T. Association rules are shaped as B =!. Rule a B B I, B I, and A is a set of things, T contains A if and only if A B implication.

Horizontal weighted association rules mining method and the improved weighted association rules mining method comparison results, wherein, shadow part is the frequent item. The algorithm of minimum support is respectively provided with 15%, 55%, 33%, to illustrate the weighting function in mining association rules [4]. Horizontal weighted association rules to generate frequent item process is first calculated by counting support degree, namely a set in the database the proportion, and then use the item weight pruning by it. As you can see from Table 1, horizontal weighted association rules mining method because of the presence of frequent item subsets is not frequent, so it can not maintain frequent item, as is shown by Equation (3).

$$\overline{I_k} = \frac{\sum_{x\in\Omega_i}[I_k(X')]}{L\cdot L}, \overline{I_{k-1}} = \frac{\sum_{x\in\Omega_i}[I_{k-1}(X)]}{L\cdot L}$$

(3)

In things set D was established, with the support of s wherein s is D things contains A!, B (i.e., A and B two) percentage, it is the probability P (B A!). Rule a B is the percentage of C, it is the conditional probability P (B A |). Namely: supp ort (A B) = P (A B!)) = P (A B D |) in things with confidence C, if D contains A affairs while Pack: x [0, 1], called the membership degree.

A collection of items called item sets (ITE set). Contains k item set called k m set. A set of frequency of occurrence is containing the item set transaction number, referred to as a set of frequency, a count or counting. If the item set is greater than or equal to the frequency of M in_sup and D in the affairs of the total number of product, the item set satisfies the small support M in_sup as is shown by Equation (4).

$$f(t) = \sum_{s=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} \widetilde{\delta f}(s,n) \psi(s, 2^s t - n)$$

(4)

The use of vertical data mining frequent pattern above two methods are based on standard data format as transaction set mining frequent patterns in the algorithm, also can be used in a vertical format of data mining [5]. It is the core of equivalence transformation algorithm. Compared with the Apfiofi algorithm uses a horizontal format data, this method has the following advantages: 1) Candidate generation process does not produce the data explosion, produced only a small number of sets of items. 2) Do not need to scan the database to determine the (K + 1) item sets (for k > = 1) of the support degree. The use of difference set technology to further reduce the storage of long TID collection overhead and the intersection computation.

Association rule mining is one of the important content of data mining field. But the traditional association rules mining is often based on such conditions: (1) The database each item of importance are the same; (2) The database is uniformly distributed: but the practical application of the database is not an ideal situation, but users on various projects for attention degree is also different, as in industry sales, some commercial profits, so the mall operators to pay close attention to it in it, for such a project should be given greater weight. The weights are introduced to the mining of association rules.

Matrix of weighted association rules mining algorithm this is built on a descending trimming weights based on 32 algorithms described in our F - R - t method based on combination weighting matrix of the rules of P g who o model inherent characteristic, out of the digging matrix of weighted association rules algorithm [6]. Proposed algorithm basic thought is: first structure F. F digging process using P tree in P tree weighting descending find frequent itemsets pruning. Then according to the matrix weighted confidence are from the frequent item sets in association rules to find exit, as is shown by Equation (5).

$$\left\| \overline{V(k,t)} \right\| = \frac{1}{n} \sum_{i=0}^{n-1} \left\| V_i(k,t) \right\|$$

(5)

When the data is updated quickly how to improve the algorithm, numerical variable processing problem, project set in the weighted case of association rule mining method of association rules mining firstly [7]. By Ag. W Imielinski and Swami proposed sweet. Apriori algorithm by A.orawal and Srikant proposed "!, subsequently it was based on the Apriori algorithm carried out a series of improvements, compared to the well-known include the use of hash table to improve the efficiency of mining association rules, uses the transaction compression technology to the scanned transaction set are compacted, used the division technology on transaction set segmentation, using sampling technology of mining and using dynamic itemset counting method.

Association rule mining has many extensions, including multilevel association rules mining, association rule mining, mining association rules based on constraints, periodic association rules mining, the mining of weighted association rules, there are some scholars of these algorithms. In addition, association rule mining technology not only can be directly used as a decision support tool, can also be applied to other data mining technology, such as association rules mining technology can be used in decision tree induction and analysis of time series data, classification in data mining technology.

3. The Research of Apriori and FP-Growth Methods

This paper analyzes the FP2 tree is applied to the maximum pattern mining, can reach a higher efficiency in 1 that the problem is mainly because the algorithm generates a large number of candidate maximum mode, testing whether they are for maximum mode needs to spend a lot of time 1 in order to solve the problem put forward the corresponding improvement measure, these methods include FP2 tree ordering, reduce generate candidate Max scale, reduce the inspection scope 1 we give the improved SFP2Max algorithm, and the algorithm with the improved algorithm and the MAF IA FP2Max before the performance comparison [8].

On the New.Apriori algorithm time complexity is too high and the number of candidate itemsets is too large, is based on the two pruning algorithm WARDM, gives the related theorem and its proof. WARDM algorithm in mining the database, only needs to scan the database once can generate all the weighted frequent itemsets, can reduce the database the number of visits, and improve the efficiency of mining. Through two times of pruning can effectively reduce the number of candidate itemsets, as is shown by Equation (6).

$$MSE(I_k, I_t) = \frac{1}{M \times N} \sum_{x=1}^{M} \sum_{y=1}^{N} \left[I_k(x, y) - I_t(x, y) \right]^2$$

(6)

FP -grow th algorithm for mining frequent pattern steps mainly is divided into two steps: FP 2tree and FP 2tree to construct a recursive mining. Construction of FP - tree 1: algorithm input: transaction database and the minimum support threshold to D m in_ sup. Output: the frequent pattern tree FP 2tree. Method: mainly consists of two steps: (1) Scans transaction database D, get the frequent 12 set F and their support count. According to F support descending sort, get frequent item table L. (2) To create the FP -tree to the root node, labeled as null " again to scan the database, in the D T ran s each transaction executes the following operation: ". Extraction of T ran s frequent item and press L order sorting. To sort after frequent item table for [P P], where p is the first element, and P is the remnant of the elements of a list. Call the in SERT tree ([P P], T).

STL provides a series of containers and template algorithm, combining these containers and algorithms can achieve a variety of utility function. At the point of chain structure will link them to have the same item _ nam e node. If the P is not empty, _ recursive invocation of in SERT tree (P, N). Analysis: FP -tree construction process of transaction database scan two times, finally the database compression storage into a tree, the tree contains frequent pattern mining all the information. FP-tree usually for long mode and intensive database with high compression ratio, but with a large number of short pattern database compression performance is superior.
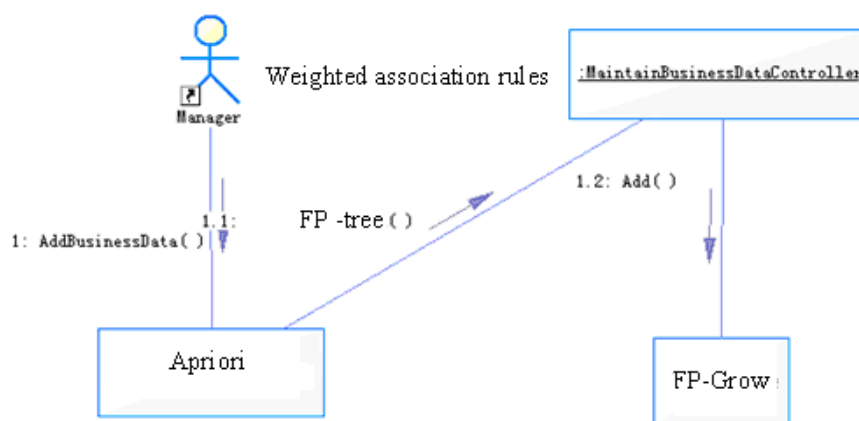

Figure 2. The Apriori and FP-Growth Methods

In the FP-Grow algorithm, the main data structure includes Item, Xiang Toubiao Header and FP 2 tree. Item header table and FP-tree element, a storage structure by defining a class to

encapsulate the related Item data members, including a item nam e _, node chain node_ link, suppo rt_ coun t support count, the parent node chain parent t_ link and CH ILD child node chain _ link, one child node chain for set container pointer types. Item header table is one by the support count in descending order linear table, data members include item name, called the _ node chain node_ link, as is shown by Figure 2.

Therefore the existence of Apriori algorithm is with the same defect. Aiming at the defects of the Apriori algorithm, the proposed a kind of not generating candidate item set algorithm of FP-Growth, to avoid the high cost of candidate itemsets generation, fewer, more efficient scanning. There is a number of FP-Growth algorithms based multi level association rule mining algorithm, the algorithm in different degree and improves the mining efficiency [9]. But most of these algorithms can only achieve a layer within the mining, unable to find different conceptual level associations among the data items; and some algorithm can realize the cross level mining, but it can not reflect the concept hierarchy of intrinsic constraints relations, resulting in the FP header between complex, can effectively simplify the process of mining.

Weighted association rules proposed by Apriori algorithm in mining frequent itemsets when there are actually two big hypothesis of the old t: (1) Of each item in the database with the same nature and function, i.e. the same. (2) The importance of each item in the database of the distribution is uniform that is the frequency of occurrence of the same or similar [10]. However, in reality world database of more than two hypotheses are not established. When the database is in project of uneven distribution of frequency difference is larger, will cause the minimum support degree is high with low have problems in two chess surface, if set too high, the mining association rules may not involve showed a lower frequency of project, as is shown by Equation (7).

$$v_X^2(\tau_j) \approx 2\int_{1/2^{j+1}}^{1/2^j} cf^\alpha df$$

$$= 2c \cdot [(2/4\tau_j)^{\alpha+1} - (1/4\tau_j)^{\alpha+1}]/(\alpha+1) \tag{7}$$

In order to overcome the shortcomings of Apriori algorithm, some scholars proposed the mining of weighted association rules algorithm, for each item were weighted, effective interest to solve the project in the database with different importance to discuss and analysis of representative weighted association rules mining algorithm named New-Apriofi algorithm.

Weighted support measure is defined, combined with this algorithm, an improved New-Apriori algorithm is put forward, in which the item set X weighted support measure is defined as: swoop (X) =max{hi, hE, hk}X ', sup (X), where hi is ij weights [11]. New, Apriori algorithm and OD algorithm of Ap. Same idea, realized by two steps : first find all weighted support degree is no less than a user specified minimum weighted support wminsup constraint all weighted frequent itemsets, denoted as L, then use a weighted frequent itemsets satisfying the minimum weighted reliability constraint of wminconf all weighted association rules.

In the mining of weighted association rules algorithm New.Apriori algorithm is the classic algorithm; it can effectively be weighted association rules mining. Based on the New.Apfiofi algorithm to produce a large number of candidate set of options and repeated scans the original database defects are improved, this paper proposes an improved algorithm WARDM algorithm. Finally through the simulation experiments validate the WARDM algorithm.

4. New Algorithms of Weighted Association Rules Based on Apriori and FP-Growth Methods

Mining frequent itemsets is the classical algorithm Apriori algorithm and FP - growth algorithm. Most other algorithms are the two algorithm variants. Classic algorithm has two default assumptions: the database of each item in the same database and the importance of each item in the distribution is uniform, the frequency of occurrence of the same or similar. Therefore, classical algorithm is equal to the consistent way of dealing with database in the database project. However the actual situation is not the case, the importance of not the same and uneven distribution. In order to solve the first problem, researchers put forward weighted association rules algorithm.

A priori algorithm there are two major premise hypotheses: 1) Database items the same importance; 2) Database of each item in the distribution is uniform, the frequency of the same or similar. However, in the real world database is often not so [12]. When the database is in the project is not uniform in distribution, frequency difference is larger, will lead to support degree of reasonable setting, if high, discovered association rules might not involve showed a lower frequency of project, as is shown by Equation (8).

$$\delta x(-1,\left[\tfrac{k}{2}\right]) = \frac{1}{\sqrt{2}}[x(0,k) - A(0)x(0,k)] = \frac{(I - A(0))}{\sqrt{2}}x(0,k)$$

(8)

The current data mining research focus. With weighting association rules mining is called weighted association rules mining. Since 1998, weighted association rule mining has received extensive attention and study, in the mining of weighted association rules algorithm research at home and abroad, a lot of results.

This paper analyzes the without generating candidate set directly to generate frequent pattern algorithm for FP 2Grow th, its basic idea is to the entire database compression represented as tree structure of FP 2tree, frequent pattern mining process into recursively generating conditions and conditions of sub Library of FP 2tree, P Rio RI performance relative to A series algorithm has a larger improve. Because FP 2Grow th algorithm data structure used to achieve more complex, has certain difficulty, and object-oriented programming language C++ the STL standard template library, wherein the data structure and algorithm for program provides powerful tools. The FP 2Grow th algorithm description and analysis, and then discusses its realization method.

In order to fully reflect the weighted association rules takes advantage of an efficient algorithm. The weighted process can not only distinguish the important degree of the mining project and the results become more reasonable, and in some cases also can greatly improve the efficiency of the algorithm, reduce mining weighted association rules required time. Because in the association rules correlation algorithm to generate frequent item sets, stage will cost algorithm's running time.

The main works are as follows: the Web data mining association rules and weighted association rule theory is studied. For a variety of weighted association rules algorithm for in-depth research, focusing on analysis of the proposed Apfiofi algorithm for the New.Apfiofi algorithm time complexity is too high and the number of candidate itemsets excessive shortage, is based on the two pruning algorithm WARDM, gives the related theorem and its proof, using synthetic data sets TIO, wIOOK and T40110D100K experiment, and the analysis of experimental results. Experimental analysis of environment: CPU Intel PE, 1GB memory, Windows XP operating system; algorithm using Java language, in Eclipse debugging development platform.
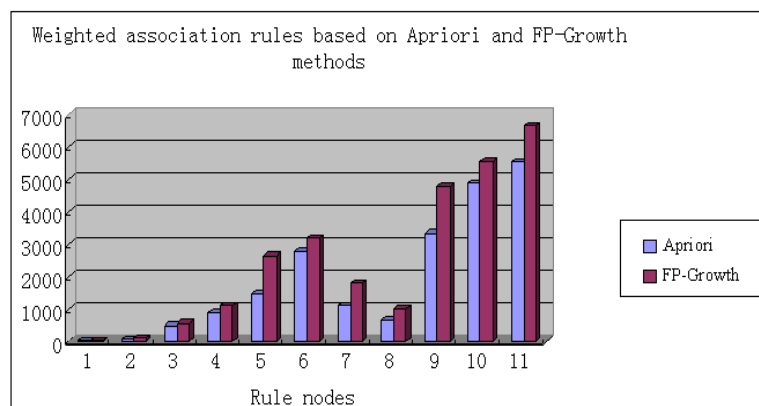

Figure 3. The Weighted Association Rules Based on Apriori and FP-Growth Methods

The paper puts forward the new algorithms of weighted association rules based on Apriori and FP-Growth methods. Algorithm in the candidate item sets Ck, Lk.1 pruning, to reduce to connect frequency (k.1). This reduces the combinatorial possibilities, which reduce circulating judgment times, thereby reducing the connecting operation, and to reduce the Ck option set quantity, improve the efficiency of the algorithm. If the large database of the time spending on data mining efficiency is very obvious.

The experiments show that the algorithm can effectively avoid single supports minimum weighted support set too low number of rules is too large and too high can not effectively the problem of mining association rules. The experimental results show that, under the same minimum weighted support threshold conditions, the WARDM algorithm of weighted frequent itemsets number less than New.Apfiofi algorithm; WARDM algorithm to generate a weighted frequent itemsets using time is less than New.Apfiofi algorithm: WARDM algorithm has good stability and expansibility, in large amounts of data mining is more high efficiency. Therefore the WARDM algorithm has better performance than New.Apfiofi algorithm.

5. Conclusion

The association rules in data mining technology especially the weighted association rules of the system, comprehensive, detailed analysis and research, and has made certain achievements. The paper puts forward the new algorithms of weighted association rules based on Apriori and FP-Growth methods. In order to improve the frequent itemsets generated layer-wise efficiency, uses the Apriori property to reduce the search space. At the same time, the algorithm keeps the A PR IO RI algorithm of excellent properties, has better time complexity and space complexity. The experimental results show that the performance of H algorithm, FP-grow faster than Apriori algorithm by an order of magnitude, although frequent pattern mining algorithm for some non dense database is very effective.

References
[1] Mohamad Farhan Mohamad Mohsin, Mohd Helmy Abd Wahab, Mohd Fairuz Zaiyadi, Cik Fazilah Hibadullah. An Investigation into Influence Factor of Student Programming Grade Using Association Rule Mining. *AISS*. 2010; 2(2): 19 ~ 27.
[2] Guo Hongjian. A Novel Frequent Item-set Mining Method based on Frequent-pattern List. *JDCTA*. 2011; 5(11): 182 ~ 188.
[3]   Xiaoke Wu, Xiongfei Li. Mining Maximal Frequent Subtrees based on Fusion Compression and FP-tree. *JDCTA*. 2011; 5(7): 191 ~ 200.
[4] Li Guodong, Xia Kewen. An Improved Data Mining Technique Combined Apriori Algorithm with Ant Colony Algorithm and its Application. *JDCTA*. 2011; 5(8): 241 ~ 249
[5] THAIR A SALIH. Performance Analysis of Chaotic Chirp Spread Spectrum System in Multipath Environment. *Journal of Theoretical and Applied Information Technology*. 2010; 18(2): 101-108.
[6] Usman Babawuro, Zou Beiji. Knowledge Representation: A General Survey and Techniques for Sound Knowledge Based Systems. *IJIIP*. 2011; 2(4): 16 ~ 22.
[7] Mingchang Liu, Hewang Liu. Research on Application of Association Rule Mining in Chinese Athletes' Nutritional and Biochemical Indexes Monitoring. *JDCTA*. 2012; 6(7): 174 ~ 180.
[8] Changjiang Li, Xianfeng Yang. The Research of Recommendation Systems in E-Commerce Based on Association Rule Improved-Apriori Algorithms. *IJACT*. 2012; 4(21): 354 ~ 361.
[9] YiJie Chen. The Development of the Commodity Flow Analysis System Based On Association Rule Mining. *IJACT*. 2012; 4(13): 430 ~ 436.
[10] CH Cai, WC Fu Ada, CH Cheng, WW Kwong. Mining association rules with weighted items. Proc. of the Int'l *Database Engineering and Applications Symposium*. 1998; 68 ~77.
[11] Charu C Agarwal, Philip S Yu. *A New Framework for Itemset Generation*. Proceedings of *the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'98)* Seattle, Washington. ACM Press Publisher. 1998; 18~24.
[12] Song YQ, Zhu YQ, Sun ZH, Chen G. An algorithm and its updating algorithm based on FP-Tree for mining maximum frequent itemsets. *Journal of Software*. 2003; 14(9): 1586-1592.