A Novel Algorithm for Mining Hybrid-Dimensional Association Rules

Chithra Ramaraju
Research Scholar
Department of Computer Applications
National Institute of Technology

Nickolas Savarimuthu
Associate Professor
Department of Computer Applications
National Institute of Technology

ABSTRACT
Association rule mining is a fundamental and vital functionality of data mining. Most of the existing real time transactional databases are multidimensional in nature. In this paper, a novel algorithm is proposed for mining hybrid-dimensional association rules which are very useful in business decision making. The proposed algorithm uses multi index structures to store necessary details like item combination, support measure and transaction IDs, which stores all frequent 1-itemsets after scanning the entire database first time. Frequent k-itemsets are generated with previous level data, without scanning the database further. Compared to traditional algorithms, this algorithm efficiently finds association rules in multidimensional datasets, by scanning the database only once, thus enhancing the process of data mining.

General Terms
Data Mining, Hybrid-dimensional association rule mining

Keywords
Multidimensional transactional databases, inter-dimensional join intra-dimensional join, Apriori algorithm, multivalued attribute, hybrid-dimensional association rules.

1. INTRODUCTION
Advancement in communication, hardware technology and sensor networks collects tremendous amount of data and subsequently stores in large number of data repositories. But the available large amount of data far exceeded human ability for comprehension, interpretation and decision making. The challenging task of efficient and effective data analysis have made promising field called data mining. Data mining is defined as "the non trivial extraction of implicit, previously unknown and potentially useful information from database". Data mining functionalities include classification, clustering, association rules, sequence mining etc. Association rule mining is one of the vital functionality for discovering interesting associations, frequent patterns, correlations, and other relationships among huge amounts of business transactional datas, with vast potential for real life applications.

Association rule mining is a two step process, namely

1. Finding all frequent itemsets
2. Generating strong association rules from frequent itemsets.

The Apriori algorithm was proposed to generate all significant frequent patterns and association rules for retail organization in the context of bar code data analysis [1]. This algorithm mines simple form of association rule called single-dimensional association rules based on Apriori property. The Apriori property states that "If any k length pattern is not frequent, its super pattern of length (k+1) is also not frequent in the database" and achieves good performance, by reducing candidate itemsets in every iteration. Number of researchers have presented many modified methods based on Apriori property. Many practical transactional databases are multidimensional in nature and some of the attributes are multivalued which poses great challenge to apply knowledge mining process. Association rules can be classified as single-dimensional association rule and multidimensional association rule based on number of predicates appearing in the rule. Multidimensional association can be classified as inter-dimensional association rule and hybrid-dimensional association rule. Hybrid-dimensional association rule involves inter-dimensional as well as intra-dimensional itemsets. Association rules generated from hybrid-dimensional itemsets have repeated predicates. In recent years, there has been lot of interest in research community for mining multilevel and multidimensional association rules. In this paper, a novel algorithm is proposed to find hybrid-dimensional association rules efficiently, without multiple scan of the database, and there is no need to check, whether to perform inter-dimensional join or intra-dimensional join between candidate itemsets. In summary, the main contribution of this works is

1. Proposing a novel algorithm with multi index structure for mining hybrid-dimensional association rules
2. Theoretical analysis of the proposed algorithm

The rest of paper is organized as follows. Section 2 summarizes some background information. Section 3 describes Apriori algorithm. Section 4 gives detailed discussion of mining multidimensional association rules and the proposed algorithm is discussed in section 5. Theoretical analysis is presented in section 6 and conclusions are given in section 7.

2. LITERATURE SURVEY

Finding frequent patterns (itemsets) play an important role in data mining and knowledge discovery techniques. Association rule describes correlation between data items in large databases or datasets. The first and foremost algorithm to find frequent pattern was presented by R. Agrawal et al. [1][2]. The Apriori algorithm finds frequent pattern of length k from the set of already generated candidate patterns of length k-1 by employing candidate generation and test methodology. This algorithm requires multiple database scans and large amount of memory to handle candidate patterns when number of potential frequent pattern is reasonably large. In the past two decades, large number of research studies have been published presenting new algorithms or extending existing algorithms to solve frequent pattern mining problem more effectively and efficiently. Most of these studies [10][13] adopts level wise candidate generation based on Apriori property. Jiawei Han et al.[8] presented FP-growth method using prefix-tree (FP-tree) for generating association rules without candidate set generation-and-test methodology.

But all the above mentioned studies are well suitable for single-dimensional transactional databases. For example, in sales transactional databases, along with items purchased, other related information like quantity purchased, price, branch location, etc. are stored. Additional related information regarding customers, customer ID, age, occupation, credit rating, income, and address are also stored in the database. Frequent itemsets along with other relevant information will be helpful in high-level decision making, which leads to challenging mining task of multilevel and multidimensional association rule mining. In recent years, there has been lot of interest in mining databases with multidimensions. Currently, many research papers have concentrated on multidimensional association rule mining and most of them are constraint based association rule mining [4][5][6][12]. Xin et al. [16] presents mining conditional hybrid-dimensional association rules, in which main attributes are marked and subordinate attributes are unmarked. Based on these marking, the algorithm performs intra-dimensional join or inter-dimensional join among itemsets. WanXin Xu et al. [15] presented a novel algorithm of mining multidimensional association rules for relational databases. In this paper, a new algorithm finding relevancy among multidimensional single valued attributes using intra-dimensional join using multi index structure, is proposed.

3. APRIORI ALGORITHM

In this section, Apriori algorithm and related basic concepts are discussed.

3.1 Association rule

Let $I = \{i_1, i_2, i_3 \ldots i_m\}$ be a set of items and D be a transaction database $D = \{T_1, T_2, T_3 \ldots T_n\}$. Each transaction $T_i \in D$ has an identifier called TID, and consists of set of items such that $T_i \subseteq I$. A is set of items and transaction T is said to contain A if and only if $A \subseteq T$.

Definition 1: Association rule is an implication of the form $A \Rightarrow B$, where A and B are itemsets, which satisfy $A \subset I$, $B \subset I$, $A \cap B = \emptyset$.

Definition 2: The association rule A=>B is true in D, with support ***s*** and confidence ***c.*** Support ***s*** is defined as , percentage of transactions in D, that contain both A and B ($A \cup B$), in transaction D. Confidence ***c*** is the percentage of transactions in D, containing A that also contains B.

$$\text{Support } (AB) = P(A \cup B)$$

$$\text{Confidence } (A \Rightarrow B) = P(B|A) = P(A \cup B) | P(A)$$

3.2 Algorithm

Apriori algorithm [1][2] employs level wise iterative approach to find all frequent itemsets. Database is scanned once to generate all frequent 1-itemset L_1 according to user specified minimum support threshold. L_1 is used to find frequent 2-itemsets L_2, by applying intra-dimensional join condition. This is repeated until no more frequent itemsets is generated. Apriori property is used to reduce number candidate itemsets in each iteration. Once all frequent itemsets are discovered, association rules are generated according to the second step in the process of association rule mining. This helps to find association and relevancy among transactional items. Apriori algorithm is aimed to find relevancy among different items of same attribute called intra-dimensional association rules. But in reality, transactional items are associated with more relevant information, which are useful for making higher level decisions. Hence hybrid-dimensional association rule mining becomes very important. It not only finds relevancy among different values of same attribute, but also finds relevancy among different values of different attributes. This type of association is called hybrid-dimensional association, which involves inter-dimensional itemsets as well as intra-dimensional itemsets. In this paper Hybrid-Dimensional-Indexing-Mining (HDIM) is proposed to generate hybrid-dimensional association rule.

4. MINING MULTIDIMENSIONAL ASSOCIATION RULES

Mining multidimensional association rule needs an enhancement to the existing algorithm or new methodology.

4.1 Multidimensional Transactional dataset

Transactional dataset D, consists of n transactions $D = \{T_1, T_2, T_3 \ldots T_n\}$. Each transaction T_i consists of m number of attributes $(d_1, d_2, d_3, \ldots d_m)$, in which d_j represents j^{th} dimension or attribute and some attributes may have multivalued categorical values. The record i can be expressed as value combination $(v_{i1}, v_{i2}, v_{i3}, v_{im})$, where v_{ij} represents i^{th} record and j^{th} dimensions, $1 \leq i \leq n$, $1 \leq j \leq m$.

4.2 Hybrid-dimensional association rules

Definition 3: Hybrid-dimensional association rule contains repeated occurrence of multi valued attributes.

Attribute of database and warehouse can be termed as predicate. Association rules are of two types. (a) Single dimensional association rules (b) Multidimensional association rules based on the number of predicates involved in the rules. In general, association rules imply single predicates called single dimensional or intra-dimensional association rules.

$$\text{Buys(X, “digital camera”)} \Longrightarrow \text{Buys(X,”HP printer”)}$$

Practical transactional database require multidimensions for storing other related information, and some attributes may be multivalued. So mining of frequent itemsets by considering other relevant information will be very useful for making decisions at higher level management like production decisions, inventory decisions.

Table 1. Sample Database

TID	A_1	A_2	A_3
1	a_{11}	a_{21}	a_{31}, a_{32}
2	a_{11}	a_{21}	a_{32}
3	a_{11}	a_{21}	a_{31}
4	a_{12}	a_{22}	a_{32}
5	a_{12}	a_{22}	a_{31}, a_{32}
6	a_{11}	a_{21}	a_{31}
7	a_{12}	a_{22}	a_{31}, a_{32}

In Table 1, attribute A_1 may represent customer age(a_{11}-young, a_{12}-middle), attribute A_2 may represent customer occupation(a_{21}-professionals, a_{22}-student) and attribute A_3 is multivalued, representing products purchased(a_{31}-computer, a_{32}-printer). Attribute values can be represented as $V_{ij(k)}$ where i^{th} record, j^{th} dimension and k^{th} value in the dimension. The first record in Table 1, is represented as

$$(v_{11\,(young)}, v_{1,2(professional)}, (v_{1,3(computer,\ printer)})).$$

Many practical databases require preprocessing process before mining hybrid- dimensional association rules. It is mandatory to have values in all dimensions of transactions and further database attribute can be categorical or quantitative. Multidimensional association rule mining uses two basic approaches to deal with quantitative attributes. The first approach uses static discretization and second uses dynamic discretization to convert quantitative attributes into categorical attributes. Association rules that involve two or more dimensions can be referred to as multidimensional association rules. Multidimensional association rule mining methods search for frequent predicates, instead of frequent itemsets. After preprocessing, it is necessary to mine association rules containing multiple predicates such as

$$\text{Age(X,”15-2”)} \wedge \text{Occupation(X,”stud”)} \Longrightarrow \text{Buys(X,”laptop”)}$$

Multidimensional association can be classified into two types.

1. Inter-dimensional association rule does not contain repeated occurrence of dimensions or predicates. For example,
   $\text{Age(X,”15-25”)} \wedge \text{Occupation(X,”stud”)} \Longrightarrow \text{Buys(X,”laptop”)}$
2. Hybrid-dimensional association rules contain repeated occurrences of some of dimensions. For example
   $\text{Age(X,”15-25”)} \wedge \text{Buys(X,”laptop”)} \Longrightarrow \text{Buys(X,”HP printer”)}.$

While generating hybrid-dimensional frequent itemsets, there could be occurrence of both inter-dimensional join as well as intra-dimensional join. Let l_1, l_2 are itemsets in L_{k-1}, the notation l_{ij} refers to j^{th} item in l_i . By convention, all items in the transactions are sorted in lexicographic order. If the attributes are single valued, inter-dimensional join is implemented. If attribute is multivalued, inter-dimensional join is implemented followed by intra-dimensional join. If the mapping is inter-dimensional join between l_1 and l_2 itemsets, it should satisfy the following condition.

$$l_1[2]=l_2[1] \wedge l_1[3]=l_2[2] \wedge \ldots \wedge l_1[k-1]=l_2[k-2] \wedge l_1[1]<l_2[k-1]$$

The items from 2^{nd} to the $(k-1)^{th}$ items of l_1 must be same as items from 1^{st} to the $(k-2)^{th}$ items of l_2. So the joining of l_1 and l_2 would result in

$$l_1[1]l_2[1]\ l_2[2]\ l_2[3]\ \ldots\ l_2[k-2]\ l_2[k-1]$$

If the mapping is intra-dimensional join between l_1, l_2, it should satisfy the following condition.

$$l_1[1]=l_2[1] \wedge l_1[2]=l_2[2] \wedge \ldots \wedge l_1[k-2]=l_2[k-2] \wedge l_1[k-1]<l_2[k-1]$$

The first (k-2) items are same in l1 and l2 and join result is

$$l_1[1]l_1[2]l_1[3] \wedge \ldots \wedge l_1[k-1]l_2[k-1].$$

Hybrid-dimensional mining is a very promising area, and has wide applications in real life. For example, In a super market, store manager may ask a question like “What group of customers would like to buy what group of items?”. In the same way, a medical officer may ask “What patient undergoing what other type of treatment?”.

4.2 Definition 4: Intra-dimensional join: An association among different values within same attributes or dimension. In Table 1, the associations between (a_{31}, a_{32}) are intra-dimensional. Only multivalued attributes uses intra-dimensional mapping.

4.3 Definition 5: Inter-dimensional join: An association among value of different attributes or dimensions. In Table 1, the association between (a_{11}, a_{21}) is inter-dimensional. Obviously all attributes uses inter-dimensional mapping.

5. HDIM (Hybrid-Dimensional Indexing Mining)

Generation of hybrid-dimensional association rule using Apriori algorithm is a time consuming process. In this section, the proposed novel algorithm HDIM is discussed. Before starting the mining process, the datasets must be preprocessed. Preprocessing includes data cleaning, integration, transformation, and data reduction and preprocessing can substantially improve the quality of mining result and time required for the mining.

5.1 Data Structure Used

The HDIM (Figure 1) algorithm defines four simple data structures namely itemsets, attribute, domain, transaction numbers respectively. These four simple structures are combined to form four level linked structure, which is used for generating (k+1) item sets. The multi-index structure is divided into two parts and first part gives attribute combination and second part provides value combination. For generating (k+1) itemsets, only previous level information is required. For the sample database, four level linked structures are shown in Figure 2. The algorithm generates frequent 1-itemsets, in the temporary table L1 along with transaction numbers, in order to compress the transaction dataset, which improves the actual time of mining. The main idea of this method is to rebuild the datasets by removing transactions which contain less than three 1-frequent itemsets. The deleted transaction numbers are removed from the temporary table L1 and IndexHead is initialized with L1. From frequent 1-itemsets, 2-itemsets are generated. Here attribute 1 is mapped with attribute 2, and 3. Attribute 2 is mapped with attribute 3. Similarly attribute 3 is mapped with itself, but there is no attribute to join. For this purpose the status of the attribute is maintained in the 1-frequent itemset. If the status is M (multivalued) , the attribute values are mapped with itself by intra-dimensional mapping, and joined with other attributes by inter-dimensional join. If the status is S, the attribute is mapped with other attributes by applying inter-dimensional join condition. From 2-itemsets, 3-itemsets are generated. The status of the attributes is required in the process of generating only 2-itemsets.While generating 3-itemsets, inter-dimensional and intra-dimensional joins are taken care of from the attribute combination. If the attribute combination is (1,2), it has to be joined with attribute which starts with 2, followed by other attribute, by using inter-dimensional join condition. If the attribute combination is (2,2) , then it has to be joined with itself using intra-dimensional join condition, and join with other attribute starting with 2 using inter-dimensional join. This is repeated until no more itemsets are generated. This structure provides all frequent itemsets starting from 1-itemsets to the longest frequent itemsets and LongHead is always pointing to the longest itemsets. But to generate (k+1) item sets, there is no need to scan the database, but the k-itemset four level linked structure is sufficient.

HDIM Algorithm:

Input: Transactionaldatabase(TDS),Min-Support(Min- sup)
Output : IndexHead (An access to all frequent itemsets)
LongHead (An access to longest itemsets)

```
Hybrid-Dimensional-Indexing-Mining(TDS, Min-sup)
{
L1 = Find-Frequent-1-Itemset(TDS);
TDS' = Trans-Compression(TDS);
IndexHead = Initialize-ItemsetSize(1);
IndexHead=Initialize-Candidate-1-Itemset(L1)
LastIndex=IndexHead;
while(L k-1 ≠ Φ )
{
  CurrIndex=Generate-Candidate-K-Itemsets(LastIndex);
  Generate-Frequent-Itemsets(CurrIndex,Min-sup)
  LastIndex → next=CurrIndex;
  LastIndex=CurrIndex;
  LongHead=CurrIndex;
}
return IndexHead;
}
Generate-Candidate-K-Itemsets(LastIndex) // k>=2)
{
CurrIndex=Initialise-Itemset-SizeNode(LastIndex-
>Itemset → size+1);
if (Attribute→ Status ='S'( for k=2)  or Attribute Combination
is Different (for  k > 2))  then
{
for each itemset l1 in  Domain of  LastIndex
for each itemset l2 in  next Attribute Domain of LastIndex
if l1[2]=l2[1] ∧ l1[3]=l2[2] ∧  .. ∧ l1[k-1]=l2[k-2] ∧  l1[1]<l2[k-1]
                                                          then
    Create a candidate K- itemset C using l1, l2
    C= l1[1]l2[1] l2[2] l2[3 ]… l2[k-2] l2[k-1]
    Insert C into CurrIndex.
    Combine -2-Itemset-to-1itemsets(CurrIndex,l1,l2)
 }
else
if ( Attribute → status ='M' (for k=2) or Attribute Combination
is same ( for k > 2))  then
{
for each itemset l1 in Domain of  LastIndex
for each itemset l2 in the same Domain  of LastIndex
l1[1]=l2[1] ∧ l1[2]=l2[2]  ∧ l1[3]=l2[3] ∧ … ∧  l1[k-1]<l2[k-1]
Create a candidate K itemset C using l1,l2
C= l1[1]l1[2]l[3]  ∧ … ∧ l1[k-1]l2[k-1]
Insert C into CurrIndex
}
return CurrIndex;
Generate-Frequent-Itemsets(CurrIndex, Min-sup)
{
 for each Itemset = (AttributePtr, DomainPtr) in CurrIndex
 if DomainPtr → Frequency >= Min-sup then
```

DomainPtr → Status = 'Yes';
else
DomainPtr → Status = 'No';
}

Figure 1. HDIM Algorithm

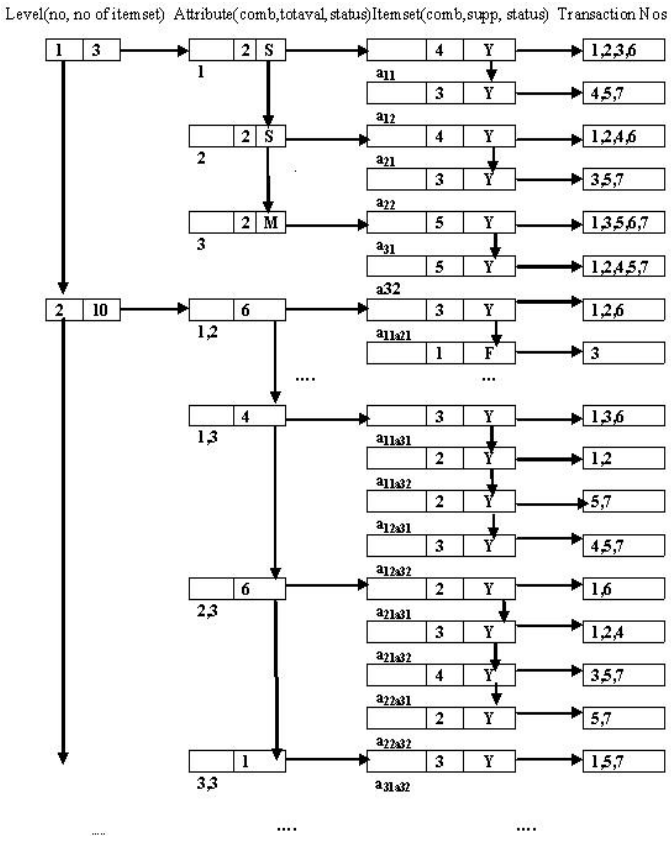

Figure 2. Four Level Index structure

6. THEORETICAL ANALYSIS

The given transactional database consists of N number of records and D number of attributes (where D << N). The cardinality of i^{th} attribute is $|V_i|$ and all the values of j^{th} dimension is $\{V_{j1}, V_{j2}, V_{j3} \ldots. V_{jp}\}$. The maximum number of items in the frequent itemset in i^{th} iteration is i. Frequent i-itemset can have $|D_i|$ different attribute combinations for inter-dimensional association, and $|V_i|$ different values of same attribute combination for intra-dimensional association. There are |Li| frequent itemsets are generated from |Ci| candidate itemsets. In this HDIM, the timing for generating frequent 1-itemsets

$$O(N*D*|Vs|) \text{ where } |Vs| = \max(|V_1|,|V_2|.... |V_d|)$$

For each value of attributes, create 4-level structure, for storing attribute values. Frequency count, status and transaction numbers are inserted in to the structure. Based on the attribute status or attribute combination, either inter-dimensional join or intra-dimensional followed by inter-dimensional join is implemented for combining two itemsets. The timing for combining two itemsets if attribute is single valued or multivalued

$$(k + (k-2)*|L_{k-1}| + |S_{(k-1),l1}| + |S_{(k-1),l2}|)$$

where $|S_{(k-1)l1}|$ is the length of transaction numbers which contain itemset l_1 and $|S_{(k-1),l2}|$ is the length of transaction numbers which contain itemset l_2.

By taking N as the maximum number of transactions, results in

$$(k + (k-2) \times |L_{k-1}| + 2N)$$

Timing for finding frequent k-itemset from candidate k-itemset is $O(C_k)$. So total time needed for HDIM algorithm is

$$O(N*D*|vs|) + \sum_{k=2}^{K} k + (k-2)*|L_{k-1}| + 2N) + O(Ck)$$

where k is negligible compared to other part, and hence time needed for HDIM algorithm is

$$O(N*D*|vs|) + \sum_{k=2}^{K} (k-2)*|L_{k-1}| + 2N) + O(Ck).$$

7.CONCLUSION

In this paper, a novel algorithm for generating hybrid-dimensional association rules is discussed. Many datasets consists of one or more multivalued attributes. By providing appropriate data structure with four level linked structures, the proposed algorithm finds hybrid-dimensional association rules efficiently from database which may have many multivalued attribute.

The strength of the algorithm is, to store the transaction numbers along with 1-itemset to avoid multiple scan of the database. Further this structure need not compare itemsets; instead it checks with attribute combination whether to proceed with inter-dimensional join or intra-dimensional join. Obviously, the comparison time is reduced to find relevancy among different values of different attributes. The algorithm can be applied for different databases, with multiple values, and performance can be studied as future work.

8. REFERENCES

[1] Agrawal, R., Imielinski, T., Swami, A., 1993. Mining Association rules between sets of items in large databases. In. Proceedings of ACM-SIGMOD, pp. 206-216.

[2] Agrawal, R. and Srikant,R. 1994. Fast algorithms for mining association rules. In Proceedings of International Conference on Very Large Data Bases (VLDB '94), pp. 487-499.

[3] Agrawal, R. and Srikant,R. 1995. Mining Sequential Patterns, In Proceedings of IEEE International Conference on Data Engineering, pp. 3-14.

[4] Anthony J.T Lee, Wan-chuen Lin, Chun-Sheng Wang , 2006. Mining association rules with multi-dimensional constraints. Elsevier, The Journal of Systems and Software 79, pp.79–92.

[5] Chuan Li, Tang, Yu, Zhang, Liu, Zhu, Jiang 2006. Mining Multi-dimensional frequent Pattern without Data Cube Construction. Springer-Verlag Berlin Heidelberg 2006, LNAI 4099, pp. 251-260.

[6] Chung-Ching Yu and Yen-Liang Chen, 2008. Mining Sequential Patterns from Multidimensional Sequence Data. IEEE Transactions on Knowledge and Data Engineering, VOL. 17, NO. 1. Pp. 136-140.

[7] Jiawei Han, Micheline Chamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, Hardcover, ISBN 1558604898.

[8] Jiawei Han, Jian Pei , Yiwen Yin, Runying Mao, Mining Frequent Patterns without Candidate Generation: A Frequent-Patterns Tree Approach, Data Mining and Knowledge Discovery,8, 53-87, 2004, Kluwer Academic Publishers .

[9] Jiawei Han , Hong Cheng, Dong Xin, Xifeng Yan, 2007. Frequent pattern mining: current status and future directions Springer Science+Business Media, LLC, Data Mining and Knowledge Discovery (2007) 15:55–86.

[10] Mannila, H., Toivonen, H., Verkamo , A.I., 1994. Efficient Algorithm for Discovering Association Rules. In Proceedings of AAAI'94 Workshop Knowledge Discovery in Databases, pp. 181-192.

[11] Ng, R., Lakshmanan , L.V.S., Han, J., Pang, A., 1998. Exploring Mining and Pruning optimization of constrained Association Rules. In Proceedings ACM-SIGMOD. Intrenational Conference on Management of Data, pp. 13-24.

[12] Runying Mao, 2001. Adaptive –FP: An efficient and Effective method for multi-level multi-dimensional Frequent pattern , Master of Science Thesis, Simon Fraser University

[13] Srikant, R., Vu, Q., and Agrawal, R. 1997. Mining association rules with item constraints. In Proc. 1997 Int. Conference on Knowledge Discovery and Data Mining, pp. 67–73.

[14] Tongyuan Wang , Huzhan Zheng, Yanjiang Qiao 2007, An Interactive Hyper Knowledge Discovery System for Chinese Medicine IEEE Fourth International Conference on Fuzzy Systems and Knowledge Discovery .

[15] WanXin Xu, RuJing Wang, 2006. A Novel Algorithm of Mining Multidimensional Association Rules. Springer-Verlag , LNCIS 344, pp. 771-60.

Yan Xin , Shi-Guang ju , 2003. Mining Conditional Hybrid-Dimensional Association Rules on the basis of Multi-dimensional Transaction Database. In Proc. Second Int. Conf. Machine Learning and Cybernetics, PP. 216-221.