Generalized Association Rule Mining Algorithms Based on Multidimensional Data

Hong Zhang and Bo Zhang

School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221008, Jiangsu, P.R. China hongzh@cumt.edu.cn

Abstract. This paper proposes a new formalized definition of generalized association rule based on Multidimensional data. The algorithms named BorderLHSs and GenerateLHSs-Rule are designed for generating generalized association rule from multi-level frequent item sets based on Multidimensional Data. Experiment shows that the algorithms proposed in this paper are more efficiency, generate less redundant rules and have good performance in flexibility, scalability and complexity.

Keywords: *Multi-dimension, Multidimensional data, Date mining, Generalized association rule, Formalization*

1. INTRODUCTION

Association rule mining is one of the most active research focuses in data mining. It was firstly proposed in the article written by Agrawal, Imielinski and Swami in 1993 [1]. And then many researchers did much hard work in association rule mining theory, algorithm design, parallel association rule mining and quantitative association rule mining. They also tried their best to improve the efficiency, adaptability and applicability of the mining algorithms and promote the application of them [1-5].

According to the limitations of the existing association rule formalization and the generalized association rule mining algorithms based on multidimensional data, this paper presents the formalized definition of generalized association rule, and designs algorithms for generating generalized association rule from multi-level frequent item sets based on multidimensional data.

2. MINING GENERALIZED ASSOCIATION RULE

2.1 The Formalization of Generalized Association Rule

The formalized description of multi-level association rule in n-dimension data set $R = (D, M, D_{str})$ is as follows:

Please use the following format when citing this chapter:

Zhang, H., Zhang, B., 2007, in IFIP International Federation for Information Processing, Volume 254, Research and Practical Issues of Enterprise Information Systems II Volume 1, eds. L. Xu, Tjoa A., Chaudhry S. (Boston: Springer), pp. 337-342.

Definition 1: an item is a 2-dimension of (d,v), item sets $I=\{(d,v)|d\in D, v\in DOM(d)\}$, where D is dimension set, $DOM(d)$ is the range of d.

Definition 2: set item $x=(d_x,v_x)\in I, y=(d_y,v_y)\in I$, if $d_x=d_y$, and $v_x\in Child(v_y)$. Then y is the father item, x is the filial item, record it as $y=\hat{x}=(d_x,\hat{v}_x)$.

Definition 3: set item sets $Z\subset I,\hat{Z}\subset I$, if Z and $\hat{Z}$ contains same items, and we can get $\hat{Z}$ by using its' father item to replace one or several items in Z, then we call $\hat{Z}$ the father item sets of Z, and Z is the filial item sets of $\hat{Z}$.

Definition 4: k-itemsets $X=\{(d_{i1},v_{i1}),(d_{i2},v_{i2}),\ldots,(d_{ik},v_{ik})\}\subset I$, where $\forall 1\le p,q\le k$, $d_{ip}\neq d_{iq}$. $\Pr(X)$ is defined as the number of transactions in original transaction database that contains the all items in item sets X. $\Pr(X)=F(d_{i1}=v_{i1},d_{i2}=v_{i2},\ldots,d_{ik}=v_{ik})$, where F refers to the function dependence relationship from the dimension set D in multi-dimension set R to measure attribute M_{count},set X is the support degree of $\sup(X)=\Pr(X)$.

Definition 5: Generalized association rule is an implication expression like $X\Rightarrow Y$, where $X\subset I$, $Y\subset I$, $X\cap Y=\phi$, and $\forall x\in X$, all $\hat{x}\notin Y$. The support degree of the rule $X\Rightarrow Y$ is $\sup(X\Rightarrow Y)=\sup(X\cup Y)$, confidence degree is *confidence* $(X\Rightarrow Y)=\sup(X\cup Y)/\sup(X)$.

2.2 Algorithms of Generalized Association Rule Mining

Descriptions of these two algorithms as follows.

2.2.1 BorderLHSs Algorithms

We can use the downward closure property based on LHS of the association rule to find the dividing line of LHSst through reverse searching means of BorderLHSs(A) under the conditiong of the given minimum support value, Description of BorderLHSs Algorithms is follows:

[Input]: Frenquent Itemset *A*

[Output]: Rule Condition (LHS) Dividing Lines (*LHSs*)

① *FIFO*={*A*}; *LHSs*=ϕ;

② while(*FIFO*$\neq\phi$) do{

③ Dequeue *B* from the head of *FIFO*;

④ onBorder=*TRUE*;

⑤ For each ($|B|$ -1)-subset C of B do {

if($P(C) \le P(A)/\min\_conf$) then {

onBorder=*FALSE*;

⑥ if (C is not in *FIFO*) then Enqueue C to the end of *FIFO*;

}

}

if (onBorder==*TRUE*) then add B to *LHSs*;

}

⑦ Answer= *LHSs*;

BorderLHSs(A) will decrease the complexity enormously, because once the item set of *LHSs* was found, the searching algorithms will stop searching other subset. Even in the worst condition, the complexity of this Algorithms is $O(2^{|A|})$.

2.2.2 GenerateRule Algorithms

GenerateRule Algorithms was obtained by deleting one frequent itemset LHSs and making it not cross with any superset or subset.

There are m frequent itemsets $A_1, A_2, \ldots, A_m$, where any itemset is the superset of (|A|+1) layer of A or subset of A. Set $B \in (\text{BorderLHSs(A)} - \bigcup_{i=1}^{m} BorderLHSs(A_i))$, relative to any other rules, $B \rightarrow (A - B)$ is irredundant.

Descriptiong of GenerateRule Algorithms is as follows:

[Input]: All Frequent Itemset L

[Output]: Irredudant Association Rule AR

① For each $A \in L$ do {

② *LHS*(A) = *BorderLHSs*(A);

③ For each $C \in L$ such that C is a $(|A|+1)$ -superset or a child itemset of A do {

LHS(A) = *LHS*(A) -*BorderLHSs*(C);

}

④ For each $B \in LHS(A)$ do {

add rule "$B \rightarrow (A - B)$" to AR;

}

}

Answer= AR;

This Algorithm gets the least and irredundant association rule. The efficiency of the association rule is improved greatly. If the processing time of every frequent itemset in set L is the same, then the computing complexity of the Algorithm and the value of set L are linearly dependent.

3. EXAMPLE VERIFICATION AND ANALYSIS

3.1 Sales Database

Given a sales database. It has four attributes: transaction identifier tid, customer's age, income and buys, where age and income are all numerical attributes,buys are category attributes, shown in table 1.

Table 1. Transaction Database of Sales

tid	age	income	buys
100	25	45k	{ IBM Laptop, HP Color Printer }
200	28	40k	{HP Desktop , Canon Color Printer }
300	44	45k	{IBM Desktop, HP Desktop }
400	21	20k	{HP Desktop, Epson b/w Printer }
500	36	40k	{ IBM Laptop }
600	32	30k	{HPLaptop, Epson b/w Printer }

3.2 Creating Association Rule

Suppose the threshold value of minimum confidence is *min_conf*=60%, according to the algorithm of GenerateLHSs-Rule, association rules are generated as shown in Table 2.

Table 2. Association Rule generated by GenerateRule (L) Algorithms

Multi-layer Association Rule	Support degree	Confidence degree
(buys, Printer) $\Rightarrow$ (age, [20,29])	3	75%
(income, [40 k,49 k]) $\Rightarrow$ (buys, Computer)	4	100%
(buys, Computer) $\Rightarrow$ (income, [40 k,49 k])	4	66.7%
(buys, Desktop) $\Rightarrow$ (age, [20,29])	2	66.7%
(age, [30,39]) $\Rightarrow$ (buys, Laptop)	2	100%
(buys, Laptop) $\Rightarrow$ (age, [30,39])	2	66.7%
(buys, Desktop) $\Rightarrow$ (income, [40 k,49 k])	2	66.7%
(buys, Laptop) $\Rightarrow$ (income, [40 k,49 k])	2	66.7%
(age, [20,29]) $\Rightarrow$ (buys, HP Desktop)	2	66.7%
(buys, HP Desktop) $\Rightarrow$ (age, [20,29])	2	66.7%
(buys, HP Desktop) $\Rightarrow$ (income, [40 k,49 k])	2	66.7%
(buys, IBM Desktop) $\Rightarrow$ (income, [40 k,49 k])	2	100%
(age, [20,29]) $\Rightarrow$ (income, [40 k,49 k]) $\wedge$ (buys, Computer)	2	66.7%

(income,[40 k,49 k]) ∧ (buys, Printer) ⇒ (age,[20,29])	2	100%
(age,[20,29]) ⇒ (income,[40 k,49 k]) ∧ (buys, Color Printer)	2	66.7%
(buys,Color Printer) ⇒ (income, [40 k,49 k]) ∧ (age [20,29])	2	100%

3.3 Interpretation of Result

- 31 association rules are generated by using general algorithms , in contrast, only 16 association rules are generated by using the algorithms in this paper. Thus, our algorithms can decrease the reduntant rules efficiently.
- Algorithms of Cumulate, Stratify and ML_T2L1 need larger store space and also have distinct limitation. However, when counting the support of itemset using the algorithms proposed in this paper, it only needs to access the relevant cell, does not need to scan the whole data cube. So it decreases the number of the candidate itemsets and improves the efficiency of generating frequent itemsets.
- Algorithm of *BorderLHSs*(A) guarantees that every subset of A can be visited one time. Once the itemset of the condition border *LHSs* was found, the searching algorithm will stop searching all the subset, which makes the complexity less than $O\left(2^{|A|}\right)$, so it can decrease the complexity of the algorithm greatly.

4. CONCLUSIONS

This paper proposes the formalized definition of generalized association rule and designs the BorderLHSs and GenerateLHSs-Rule Algorithms of generalized association rule mining based on multidimensional data, which decrease the number of redundant rules efficiently. Experiment shows that the algorithms in this paper is superior on algorithms efficiency and generating irredundant rules. At the same time, the algorithms have good performance in flexibility, scalablicity and complexity. This paper also has great theorical meaning and practical value on generalized association rule mining based on multidimensional data.

This paper is supported by the Natural Science Foundation of Jiangsu Province (serial number: BK2005021).

REFERENCES

1. R. Agrawal, T. Imielinski, and A. Swami, Mining association rules between sets of items in large databases, in *Proc. of ACM SIGMOD Conference on Management of Data*, eds. P. Buneman and S. Jajodia (ACM Press: Washington D.C, 1993), pp.207-216.
2. G. Piatetsky-Shapiro and W.J. Frawley, *Knowledge Discovery in Databases* (AAAI/MIT Press: Menlo Park, California, 1991).
3. J. Han and M. Kamber, *Data Mining——Concept and Technology* (China Machine Press: Beijing, 2001).
4. J. Li and H. Gao, Multidimensional Data Modeling for Data Warehouses, *Journal of Software*. Volume 11, Number 7, pp.908-917, (2000).

5. M. Chen, J. Han, and P.S. Yu, Data mining: An Overview from a Database Perspective, *IEEE Trans. on Knowledge and Data Engineering*. Volume 8, Number 7, pp.866-883, (1996).