Association Rule and Quantitative Association Rule Mining among Infrequent Items

Ling Zhou、Stephen Yau

Multimedia Data Mining'07, August 12, 2007

M9690214 蕭惠文

INTRODUCTION

- The main goal of association rule mining is to discover relationships among set of items in a transactional database.
- An association rule is an implication of the form $A \Rightarrow B$, where A and B are frequent itemsets in a transaction database and $A \cap B = \emptyset$.
- There is an increasing demand of mining the infrequent items (such as rare but expensive items).

- In this paper, Matrix-Based Scheme and Hash-Based Scheme to explore interesting among infrequent items.

BASIC CONCEPTS

- Let A, B be a set of items, an association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cap B = \emptyset$.
- The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain both A and B.
  - the support of the rule is the probability $P(A \cup B)$.
- The rule $A \Rightarrow B$ also has another measure called confidence c where c is the percentage of transactions in D containing A that also contain B.
  - the confidence of the rule is the conditional probability $P(B|A)$.
- The rules that have a support and confidence greater than predefined thresholds are called valid (or strong) rules.

PRUNING STRATEGIES (1-2)

- Only consider infrequent itemsets, which contain infrequent items.

- X and Y are independent if $P(X \cup Y) = P(X)\ P(Y)$.
- So rule $X \Rightarrow Y$ is not interesting if $supp(X \cup Y) \approx supp(X) * supp(Y)$, which means that a rule is not interesting

  if its antecedent and consequent are approximately independent.

- $interest(X, Y) = |supp(X \cup Y) - supp(X)supp(Y)|$.
  - If $interest(X, Y) \geq min\_interest$ (predefined threshold)
  - itemset $X \cup Y$ is referred to as a potentially interesting itemset.

PRUNING STRATEGIES (2-2)

Definition 1

- I is an infrequent itemset of potential interest if:
  - $\exists X, Y: X \cap Y = \emptyset, X \cup Y = I$,
  - for $\forall ik \in X, jk \in Y$,
    $supp(ik) , supp(jk) \leq min\_support$
  - $interest(X, Y) \geq min\_interest$.

CORRELATION ANALYSIS (1-2)

- if $P(X \cup Y) = P(X)\ P(Y)$
- otherwise itemsets X and Y are dependent and correlated as events.
- The correlation (dependence) between the occurrence of X and Y can be measured by correlation(X,Y)

$$correlation(X,Y) = \frac{P(X \cup Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{P(Y)} = \frac{P(X \mid Y)}{P(X)}$$

  - If correlation(X, Y) = 1 or P(Y|X) = P(Y)[or P(X|Y) = P(X)]
  - If correlation(X, Y) >1 or P (Y|X) >P(Y) [or P (X|Y) >P(X)]
  - If correlation(X, Y)< 1or P (Y|X)< P(Y) [or P (X|Y) <P(X)]

CORRELATION ANALYSIS (2-2)

- CPIR(Y |X) as the confidence measure of an association rule between itemsets X and Y.
- The confidence measure of rule X ⇒Y is defined as

confidence(X ⇒ Y) = CPIR(Y |X)

= [P(Y |X) – P(Y)]/[1 – P(Y)]

=[P (X ∪ Y)– P(X)P(Y)]/[P(X)(1 – P(Y))]

=[supp(X ∪ Y) – supp(X)supp(Y)]/[supp(X)(1 – supp(Y))]

- If P(Y |X) = P(Y), X ⇒ Y, confidence(X ⇒Y) = CPIR(Y |X) = 0
- If P(Y |X) > P(Y), X⇒Y, confidence(X⇒Y) =CPIR(Y |X) = 1
- If P(Y |X) < P(Y), X⇒Y, confidence(X ⇒ Y) =CPIR(Y |X) = -1

DISCOVERING ASSOCIATION RULES AMONG INFREQUENT ITEMS (1-7)

Association Rules of Interest among Infrequent Items

Definition 2

- I be the set of items in a database D,
- $J = A \cup B$ be an itemset , $A \cap B = \emptyset$, $supp(A) \neq 0$, $supp(B) \neq 0$,
- threshold min_support, min_confidence and min_interest > 0
  - if $supp(A), \ supp(B) \leq min\_support$,
  - $interest(A,B) \geq min\_interest$, $correlation(A,B) > 1$
  - $CPIR(B|A) \geq min\_confidence$,

- $A \Rightarrow B$ is a rule of interest

DISCOVERING ASSOCIATION RULES AMONG INFREQUENT ITEMS (2-7)

Association Rule Mining Process among Infrequent Items

- **Phase 1.** Identify all infrequent itemsets of potential interest
- **Phase 2.** Extract rules of interest from these itemsets

DISCOVERING ASSOCIATION RULES AMONG INFREQUENT ITEMS (3-7)

Matrix-Based Scheme (MBS)

- there are 5 transactions and 6 items.
- min_support=50%.

Table 1: Purchase of *computer game* and *video* in an electronic store

TID	ID of Items
T1	I2, I1, I0, I5
T2	I3, I1, I4
T3	I4, I3
T4	I2, I1
T5	I4, I0, I1

DISCOVERING ASSOCIATION RULES AMONG INFREQUENT ITEMS (4-7)

Matrix-Based Scheme (MBS)

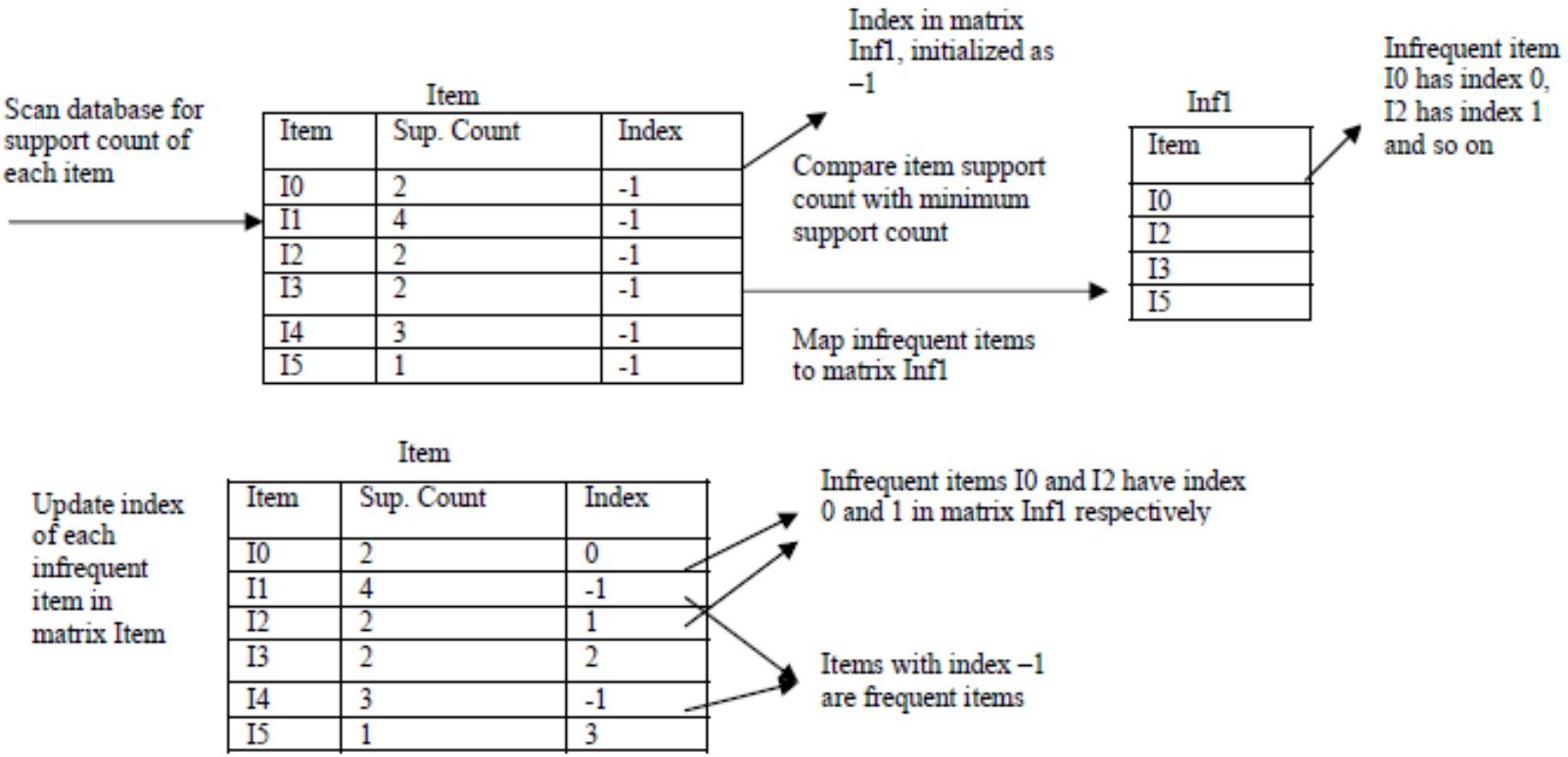


Figure 1: Identification of infrequent items in database D

DISCOVERING ASSOCIATION RULES AMONG INFREQUENT ITEMS (5-7)

Matrix-Based Scheme (MBS)

Inf2

2-itemsets	{I0, I2}	{I0, I3}	{I0, I5}	{I2, I3}	{I2, I5}	{I3, I5}
Sup. count	0	0	0	0	0	0

Inf3

3-itemsets	{I0,I2,I3}	{I0,I2,I5}	{I0,I3,I5}	{I2,I3,I5}
Sup. count	0	0	0	0

Figure 2: Matrixes to store support counts of infrequent k-itemsets **Infk(k=2,3…)**

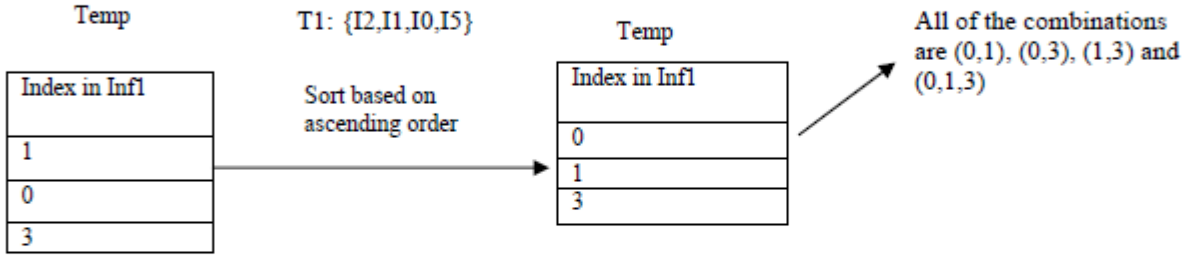


Figure 3: Temporary matrix to store index values of infrequent items in each

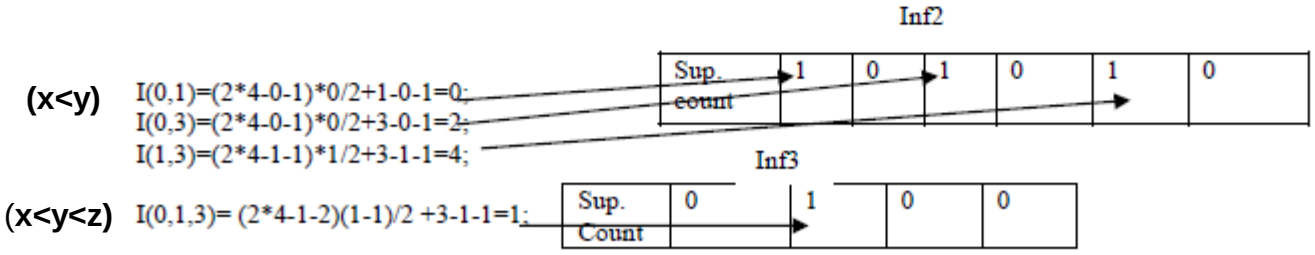


Figure 4: Updated support counts of k-itemsets

DISCOVERING ASSOCIATION RULES AMONG INFREQUENT ITEMS (6-7)

Matrix-Based Scheme (MBS)

Infrequent 2-itemsets

Itemsets	Sup. Count
{I0,I2}	1
{I0,I5}	1
{I2,I5}	1

Infrequent 2-itemsets of interest

Itemsets	Sup. Count
{I0,I5}	1
{I2,I5}	1

Figure 5: Generation of infrequent k-itemsets of interest

- Compute value of correlation(X,Y), if correlation(X,Y)>1, then employ interestingness measure, CRIP(X,Y), to extract rules of strong interest.

DISCOVERING ASSOCIATION RULES AMONG INFREQUENT ITEMS (7-7)

Hash-based Scheme (HBS)

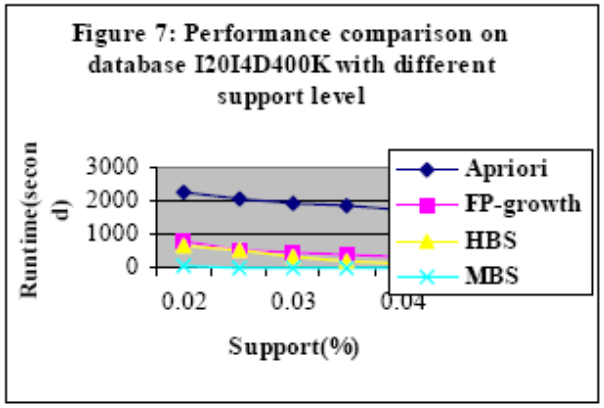
Create hash table using hash function h(x, y)

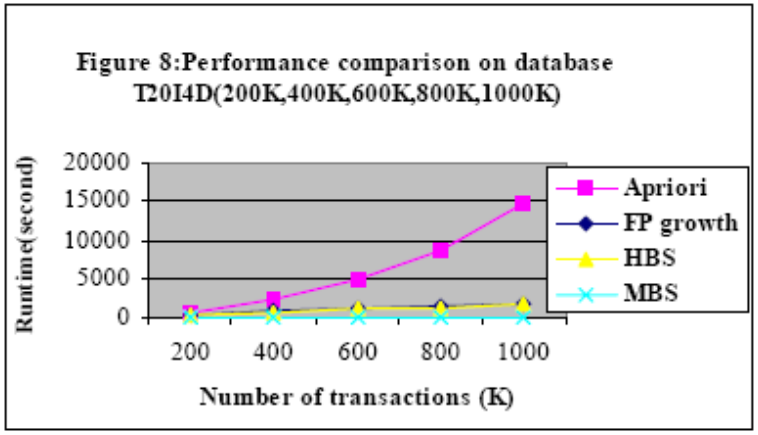
Bucket address	0	1	2	3	4	5	6
Bucket count	2	1	1	2	1	1	1
Bucket contents	{I0,I1} {I0,I1}	{I0,I2}	{I0,I5}	{I1,I2} {I1,I2}	{I1,I5}	{I2,I5}	{I0,I2,I5}

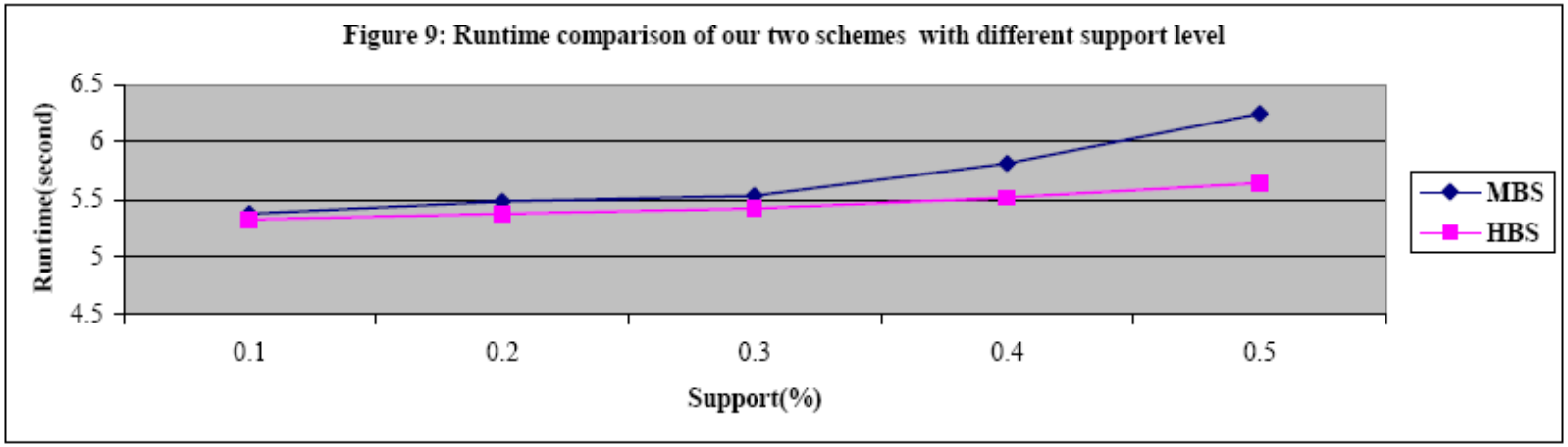
Figure 6:Hash table generated to store support counts of all infrequent k-itemsets

- Use function, interest(X,Y), to prune uninteresting itemsets.
- Employ functions, correlation(X,Y) and CPIR(X,Y), to capture rules of strong interest.

EXPERIMENTAL EVALUATION AND PERFORMANCE STUDY







QUANTITATIVE ASSOCIATION RULE MINING AMONG INFREQUENT ITEMS (1-2)

- several interesting rules from MBS (or HBS), which are {necklace⇒earring}, {table⇒chair} and {water⇒beverage}
- support of 10% and confidence of 60%
- there are total 10 transactions containing these three itemsets
- support of Qitemset based on the total number of transactions containing that itemset. Qminsup, be 30%.

QUANTITATIVE ASSOCIATION RULE MINING AMONG INFREQUENT ITEMS (2-2)

Simple rule

Table 5: Itemset {necklace, earring} with its quantitative itemsets

Itemset/Qitemset	# of transactions
{necklace, earring}	10
{necklace=1, earring=2}	6
{necklace=1, earring=1}	2
{necklace=1, earring=3}	2

General rule

Table 6: Itemset {table, chair} with its quantitative itemsets

Itemset/Qitemset	# of transactions
{table, chair}	10
{table=1, chair=3}	1
{table=1, chair=4}	3
{table=1, chair=5}	3
{table=1, chair=6}	3

Semantic rule

Table 7: Itemset {water, beverage} with its quantitative itemsets

Itemset/Qitemset	# of transactions
{water, beverage}	10
{water=1, beverage=8}	4
{water=4, beverage=4}	1
{water=7, beverage=2}	4
{water=5, beverage=4}	1

"large quantity"→ if n(A) ≥7
"medium quantity" →if 3≤n(A)<7
"small quantity" →if n(A)<3.

CONCLUSIONS AND FUTURE WORK

- In this paper, we propose two novel algorithms called MBS and HBS for effcient discovery of association rules among infrequent items.

- To further develop our research,the idea of using constraints can further help reduce the size of itemsets generated.