Multilevel Association Rules By Mining both Positive and Negative Approach

May Thu Htun
University of Computer Studies (Mandalay)
extrapowerfish@gmail.com

Abstract

Association rule mining is one of the key issues in knowledge discovery. The discovery of frequent patterns, association, and correlation relationship among huge amounts of data is useful in selective marketing, decision analysis and business management. Association rules are traditionally defined as implications of the form A=>B, where A and B are frequent itemsets in a transaction database. The method extends traditional associations to include association rules of forms A => ¬B, ¬ A => B, and ¬ A => ¬ B, which indicate negative associations between itemsets. The negative rules are generated from infrequent itemsets. This system generates the set of frequent itemsets and the set of infrequent itemsets with three database including books, electronic and grocery store of sale transaction. This system presents a method for mining both positive and negative association rules. This system demonstrates that experimental results and efficiency of both positive and negative association rules.

1. Introduction

The discovery of association relationships among huge amounts of data is useful in selective marketing, decision analysis and business management. A popular area of application is market basket analysis, which studies the buying habits of customers by searching for sets of items that are frequently purchased together [3].

Association rules are traditionally defined as implications of the form A=>B, whose support (supp) and confidence (conf) meet some system specified minimum support (minsupp) and minimum confidence (minconf) thresholds respectively[4]. A and B here are disjoint itemsets (A ∩ B = 0). This is the support-confidence framework for association analysis that was first proposed by Agrawal, Imielinski, and Swami (Agrawal, Imielinski, and Swami 1993). In practical applications, the rule A A=>B can be used to predict that "if A occurs in a transaction, then B will likely also occur in the same transaction". Such applications are expected to increase product sales and provide more convenience for supermarket customers. This paper extends this definition to include association rules of forms *A => ¬B, ¬ A => B, and ¬ A => ¬ B*, which indicate negative associations between itemsets. The rule of the form A=>B is positive rules, and the rules of the other forms are negative rules, Negative rules are also very useful in association analysis, although they are hidden and different from positive rules. This paper presents a method that mines both positive and negative rules. The positive rules are generated from frequent itemsets. The negative rules are generated from infrequent itemsets [2].

The rest of the paper is structured as follows: Section 2 discusses the related work. Then, Section 3 presents the overview of the background theory and Section 4 show system flow diagram and Section 5 describes Implementation and Section 6 describe Experimental result , Section 7 show Efficiency , Section 8 describe Conclusion.

2. Related Work

The Chi-square test based model in (Brin, Motwani and Silverstein 1997) first mentioned negative relationships between two frequent itemsets. The Chi-square value for itemsets X and Y can be used to determine whether X and Y are independent, a metric is needed to determine whether the correlation between X and Y is positive or negative. However, if the correlation is negative, other methords must be applied to determine which of X⇒¬Y, ¬X⇒Y and ¬X⇒¬Y can be extracted as a valid rule and to compute the support, confidence and interest for such a rule.Therefore this modle has not addressed how to mine negative association rules [2].

(Savasere,Omiecinski and Navathe 1998) addresses the issues of negative rule mining called strong negative association mining. Previously discovered positive associations are combined with domain knowledge in the form of taxonomy for mining association rules. This model is knowledge-dependent and can discover negative associations of the form A ⇒¬B. However, it is not clear in this model which one of $A \Rightarrow \neg B$, $\neg A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ is the actual relationship between A and B. This paper does not require domain knowledge, negative association rules are given in more concrete expressions to indicate actual relationships between different itemsets and we have designed a general framework for mining both positive and negative association rules at the same time.

3. Frequent and Infrequent Itemsets

A frequent itemset (also called large itemset (Chan, Han, and Yu 1996) is an itemset that meets the system-specified minimum support. An infrequent itemset (or small itemset) as an itemset that does not meet the system-specified minimum support [2].

3.1 Frequent Itemsets for Positive Association Rules

Let X, Y $\subseteq$ I [5] be two itemsets, $X \cap Y = \emptyset$, $supp(X) \neq 0$, $supp(Y) \neq 0$, and the minimum support (minsupp) and minimum confidence (minconf) be given by the system. $X \Rightarrow Y$ can be extracted as a valid rule if

(1) $Supp(X \cup Y) = p(X \cap Y) \geq minsupp$, and

(2) $conf(X \Rightarrow Y) = p(Y/X) \geq minconf$.

A rule $X \Rightarrow Y$ is not interesting if
$support(X \cup Y) \approx support(X) \times support(Y)$.

This argument proposed that only if $supp(X \cup Y) - supp(X)\ supp(Y) \geq$ minimum interest, the rule $X \Rightarrow Y$ is of interest. $X \Rightarrow Y$ is a valid positive rule of interest if and only if –

(1) $X \cap Y = \emptyset$

(2) $supp(X \cup Y) \geq minsupp$

(3) $supp(X \cup Y) - supp(X)supp(Y) \geq$ minimum interest

(4) $supp(X \cup Y)/supp(X) \geq minconf$,

Where mininterest is a minimum interest specified by the system, and $X \cup Y$ is frequent itemset [2].

3.2 Infrequent Itemsets of Interest for Negative Association Rules

To mine negative association rules, all itemsets for possible negative association rules in a given database need to be considered. For example, if $A \Rightarrow \neg B$ can be discovered as a valid rule, then $supp(A \cup \neg B) \geq$ minimum support must hold. If minimum support is high, $supp(A \cup \neg B) \geq$ minimum support would mean that $supp(A \cup B) <$ minimum support and itemset $A \cup B$ cannot be generated as a frequent itemsets in existing association analysis algorithms. In other words, $A \cup B$ is an infrequent itemsets. However, there are too many infrequent itemsets in databases, and we must define some conditions for identifying itemsets of interest. If A is a frequent itemsets and B is an infrequent itemset with frequency 1 in a large database, then $A \Rightarrow \neg B$ certainly looks like a valid negative rule, because $supp(A) \geq$ minimum support, $supp(B) \approx 0$, $supp(A \cup \neg B) \approx supp(A) \geq$ minimum support, $conf(A \Rightarrow \neg B) = supp(A \cup \neg B)/supp(A) \approx 1 \geq minconf$. This could indicate that the rule $A \Rightarrow \neg B$ is valid, and the number of this type of itemsets in a given database can be very large. This means that if $A \Rightarrow \neg B$(or $\neg A \Rightarrow B$ or $\neg A \Rightarrow \neg B$) is a negative rule of interest, A and B would be frequent itemsets. In other words, no matter whether association rules are positive or negative, we are only interested in frequent itemsets that occur in association rules. They take this observation as one of the main conditions for identifying infrequent itemsets for mining negative association rules [1].

This paper defines the following conditions for a rule $A \Rightarrow \neg B$ to be a valid negative rule of interest.

(1) $A \cap B = \emptyset$

(2) $supp(A) \geq$ minimum support, $supp(B) \geq$ minimum support, and $supp(A \cup \neg B) \geq$ minimum support,

(3) $supp(A \cup \neg B)$ - $supp(A)supp(\neg B) \geq$ minimum interest,

(4) $supp(A \cup \neg B)/supp(A) \geq minconf$.

This paper can define conditions for rules of the forms $\neg A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ accordingly. When $A \Rightarrow \leftarrow B$ is a valid negative rule, $A \cup B$ is an infrequent itemsets of interest. If i is an infrequent itemset of interest, there is at least one expression $i = A \cup B$ such that one of the rules $A \Rightarrow \neg B$., $\neg A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ is a valid negative association rule of interest.

3.3 Identifying Frequent and Infrequent Itemsets of Interest

All Itemsets of Interest generates all frequent and infrequent itemsets of interest in a given database D, where PL is a the set of frequent itemsets of interest in D and NL is the set of all infrequent itemsets of interest in D. PL and NL contain only frequent and infrequent itemsets of interest respectively, and all frequent itemsets in Frequent i ($i > 0$) must be saved for generating future infrequent itemsets of interest.

Algorithm All Itemsets of Interest finds all interesting frequent and infrequent itemsets of interest. Since the frequent itemsets are generated in the same way in Apriori, All Itemsets of Interest can generate all frequent itemsets. The pruning is verifies the minimum interest requirement, and improves Apriori by avoiding generating uninteresting frequent itemsets [2].

Procedure: All Itemsets of Interest

Input: D – a database; minsupp – minimum support;

mininterest – minimum interest;

Output: PL – set of frequent itemsets of interest ; NL –set of infrequent itemsets of interests;

(1) Let PL ← Ø; NL ← Ø;

(2) Let $Frequent_1$ ← BEGIN $frequent_1$-itemsetsEND; PL ← PL U Frequent1;

Let L_1 ← $Frequent_1$; S_1 ← Ø

(3)For (K=2; L_{k-1} ≠ Ø; k++) do

begin // Generate all possible frequent and infrequent

k-itemsets of interest in D.

(3.1) let Tem_k ← the k –itemsets constructed from $Frequent_{k-1}$;

(3.2) for each transaction t in D do

// Check which k- itemsets are included in t.

begin

let Tem_t ← k- itemsets in both t and Temk;

for each itemset A in Tem_t do

let A.count ← A.count+1;

end

(3.3) let $Frequent_k$ ← BEGINc|c ∈ Tem_k Λ (supp(c) =

(c.count/ |D|) >= minsupp) END;

let L_k ← $Frequent_k$;

let S_k ← Tem_k – $Frequent_k$;

(3.4)// Prune all uninterested k-itemsets in L_k

for each itemset i in L_k do

if i is uninterested by the mininterest

then let L_k ← L_k – BEGINiEND;

let PL ←PL U L_k;

(3.5) // Prune all uninterested k-itemsets in S_k

for each itemset i in S_k do

if i is uninterested by the mininterest

then let S_k ← S_k – BEGINiEND;

let NL ← NL U S_k;

end

The initialization is done in Step (1). Step (2) generates Frequent1 of all frequent 1-itemsets in database D in the first pass of D. Step (3) generates Lk and Sk for k ≥ 2 by a loop, where Lk is the set of all frequent k- itemsets of interest in the kth pass of D, Sk is set of all infrequent k-itemsets of interest, and the end-condition of the loop is Lk-1 = Ø. For each pass of the database in Step (3), say pass k, there are five substeps as follows.

Step (3.1) generates Temk of all k-itemsets in D, where each k-itemset in Temk is the union of two frequent itemset5s in Frequentk-1. .Each itemset in Temk is counted in D by a loop in Step (3.2). Frequentk, Lk and Sk are generated in Step (3.3). Both Frequentk and Lk are sets of all frequent k-itemsets in Temk that meet minsupp. That is , Lk is the set of all frequent k-itemsets in Temk, Sk is the set of all infrequent k-itemsets in Temk, whose supports do not meet minsupp, and Sk = Temk – Frequentk. Sk is the set of all possible infrequent k-itemsets in Temk.

3. Extracting Positive and Negative Association Rules

Algorithm Positive and Negative Associations generates not only all positive association rules in frequent itemsets of interest but also negative association rules in infrequent itemsets of interest.

3.4 Four Types of Association Rules

Let I be the set of items in a database D, i= A∪B⊆ I be an itemsets, A∩B = Ø, supp(A) ≠ 0, supp(B) ≠ 0, and the minimum support (minsupp) and minimum confidence (minconf) and minimum interest > 0 be given by the system. Then,

1. If supp (A∪B) ≥ minimum support, supp (A∪B) - supp (A) supp (B) ≥minimum interest and PR(B/A) ≥ minconf, then A⇒B can be extracted as a positive rule of interest.
2. If supp(A∪¬B) ≥ minimum support, supp(A) ≥ minimum support, supp(B) ≥ minimum support, supp(A∪¬B) - supp(A)supp(¬B) ≥ minimum interest and PR(¬B/A) ≥ minconf, then A ⇒ ¬B can extracted as a negative rule of interest.
3. If supp(¬A∪B) ≥ minimum support, supp(A) ≥ minimum support, supp(B) ≥ minimum support, supp(¬A∪B) - supp(¬A)supp(B) ≥ minimum interest and PR(B/¬A) ≥ minconf, then ¬A ⇒ B can extracted as a negative rule of interest.
4. If supp(¬A∪¬B) ≥ minimum support, supp(A) ≥ minimum support, supp(B) ≥ minimum support, supp(¬A∪¬B) - supp(¬A)supp(¬B) ≥ minimum interest and PR(¬B/¬A) ≥ minconf, then ¬A ⇒ ¬B can extracted as a negative rule of interest [2].

3.5 Algorithm Design

Input: D - a database; minsupp, minconf, mininterest-threshold values;

Output: association rules;

(1) Call procedure All Itemsets of Interest;

(2) // Generate positive association rules in PL.

for each frequent itemset A in PL do

For each itemset X U Y = A and X ∩ Y =Ø do

begin

if supp (X U Y) – supp (X) supp(Y) ≥ mininterest

eq (1)
then
if $PR(Y|X) \geq minconf$ then eq (2)
output the rule $X \Rightarrow Y$;
if $PR(X|Y) \geq minconf$ then
output the rule $Y \Rightarrow X$;
end;
(3) //Generate all negative association rules in NL.
for each itemset A in NL do
for any $X \cup Y = A$ and $X \cap Y = \emptyset$ do
begin
(3.1) // Generate negative association rules of the forms $\neg X \Rightarrow Y$ and $Y \Rightarrow \neg X$.
If $supp(X) \geq minsupp$ and $supp(Y) \geq minsupp$ and $supp(\neg X \cup Y) \geq minsupp$ then eq (3)
if $supp(\neg X \cup Y) - supp(\neg X)\,supp(Y) \geq mininterest$ then eq (4)
begin
if $PR(Y|\neg X)| \geq minconf$ then eq (5)
output the rule $\neg X \Rightarrow Y$;
if $PR(\neg X|Y)| \geq minconf$ then
output the rule $Y \Rightarrow \neg X$;
end;
(3.2) // Generate negative association rules of the forms $\neg X \Rightarrow \neg Y$ and $\neg Y \Rightarrow \neg X$
if $supp(X) \geq minsupp$ and $supp(Y) \geq minsupp$ and $supp(\neg X \cup \neg Y) \geq minsupp$ then eq (6)
if $supp(\neg X \cup \neg Y) - supp(\neg X)\,supp(Y) \geq mininterest$ then eq (7)
begin
if $PR(\neg Y|\neg X)| \geq minconf$ then eq (8)
output the rules $\neg X \Rightarrow \neg Y$;
if $PR(\neg X|\neg Y)| \geq minconf$ then
output the rules $\neg Y \Rightarrow \neg X$;
end; end;
(4) return.

3.6 Example Evaluation

Let for input, minsupp = 0.3 , minconf = 0.5 , mininterest = 0.05 BD is a frequent 2 itemset of interest in PL and supp(B)=0.7 and supp(D)=0.6 in frequent itemset in database.
For itemset B U D in PL
By Algorithm,
if supp (B U D) – supp (B) supp (D) >= mininterest eq (1)

$$0.6 - (0.7 * 0.6) >= 0.05$$
$$0.18 >= 0.05 \text{ then}$$

if PR (D|B) >= minconf eq (2)

$$PR(D|B) = \frac{supp(D \cup B) - supp(B)\,supp(D)}{supp(B)\,(1 - supp(D))}$$

$$= \frac{0.6 - (0.7 * 0.6)}{0.7 * (1 - 0.6)}$$

$$= 0.643 >= 0.5 \text{ then}$$

Output the rule B => D
BE is a infrequent 2 itemset of interest in NL.and supp(B)=0.7 supp(E)=0.3 in infrequent itemset in database.
For itemset B U E in NL
By Algoritm,
if supp (B) = 0.7 >= minsupp=0.3 and supp (E)=0.3 >= minsupp=0.3 and Supp (B U ¬ E) =0.6 >= minsupp=0.3 then
if **supp**(B U ¬ E) – supp (B) supp (supp ¬ E) >= mininterest eq (4)

$$0.6 - (0.7 * 0.7) >= 0.05$$
$$0.11 >= 0.05 \text{ then}$$

if PR (¬ E| B) >= minconf eq (5)
PR (¬ E | B)

$$= \frac{supp(B \cup \neg E) - supp(B)\,supp(\neg E)}{supp(B)\,(1 - supp(\neg E)}$$

$$= \frac{0.6 - (0.7 * 0.7)}{0.7 * (1 - 0.7)}$$

$$= 0.524 >= 0.05 \text{ then}$$

Output the rule B >= ¬ E

4. System Flow Diagram

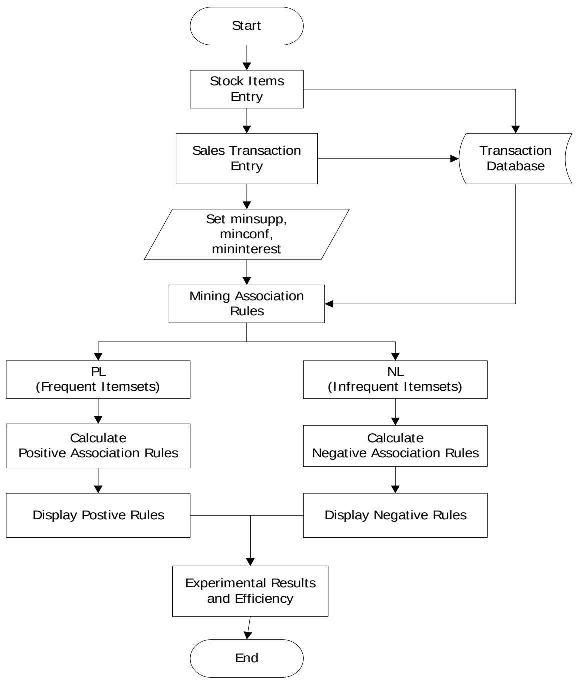

Figure 1 System flow diagram of the system

In Figure 1, the system imports databases and then the system inserts minmum support (minsupp), minimum confidence (minconf) and minimum interest (mininterest). Then, the system generates association rules. The system also generates frequent itemsets and infrequent itemsets from association rules. The system calculates positive association rules from frequent itemsets and negative association rules from infrequent itemsets.

The system also views experimental results of number of positive rules and negative rules. This system shows the run time comparison of MBP (Mining By Pruning) and MNP (Mining with No-Pruning). So the system can view the run time comparison of the system.

5. Implementation

This paper generates the association positive rules and negative rules from positive and negative itemsets by using previous algorithms (presented in Section 2.3 and Section 3.2).

5.1 Frequent and Infrequent Itemsets

In Figure 2, generate set of frequent itemsets of interest for positive rules and set of infrequent itemsets of interest for negative rules.

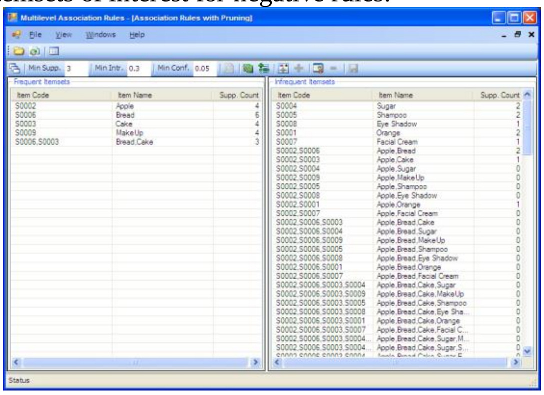

Figure 2 Generating Frequent and Infrequent Itemsets Form

5.2 Positive Association Rules Form with Pruning

After generating frequent itemsets and infrequent itemsets, extract positive rules of the form $A \Rightarrow B$ in frequent itemsets. This system shows positive rules by Items and by Codes and their conf in Figure 3.

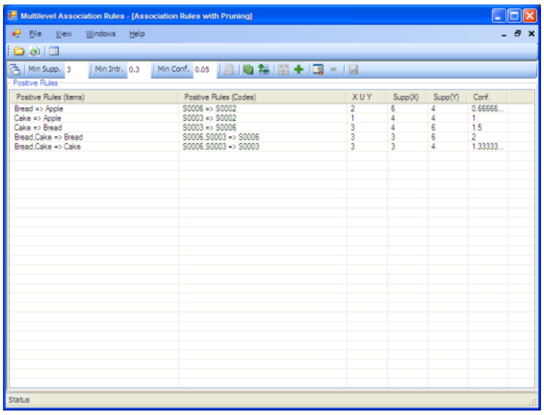

Figure 3 Generating Positive Association Rules Form with Pruning

5.3 Negative Association Rules Form with Pruning

Extract negative rules of the forms $A \Rightarrow \neg B$, $\neg A \Rightarrow B$, $\neg A \Rightarrow \neg B$ in infrequent itemsets. In this form, the system can view negative rules by Items and by Codes and their conf in Figure 4.

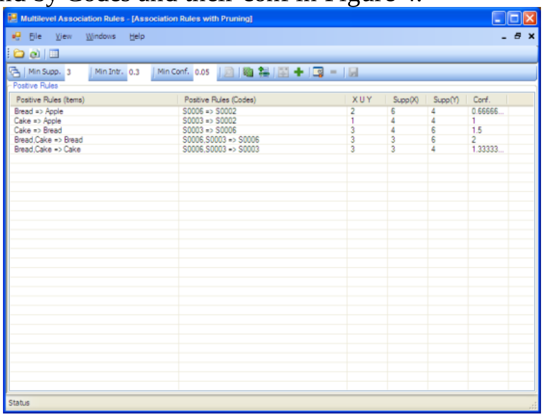

Figure 4 Generating Negative Association Rules Form with Pruning

5.4 Positive Association Rules Form without Pruning

Extract positive rules of the form $A \Rightarrow B$ in frequent itemsets without generating frequent itemsets and infrequent itemsets from association rules in Figure 5.

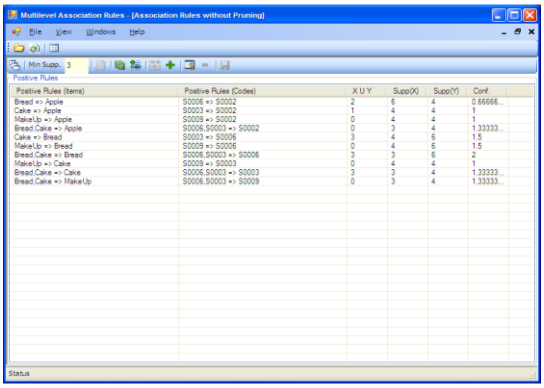

Figure 5 Generating Positive Association Rules Form without Pruning

5.5 Negative Association Rules Form without Pruning

Extract negative rules of the form A ⇒ B in frequent itemsets without generating frequent itemsets and infrequent itemsets in Figure 6.

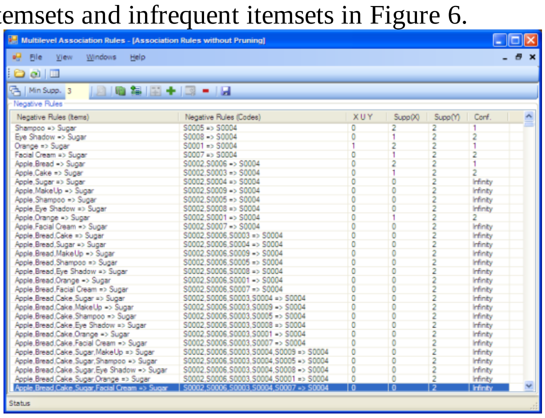

Figure 6 Generating NegativeAssociation Rules Form without Pruning

6. Experimental Results

This paper demonstrates that experimental results of how many positive rules and negative rules in Table 1. This paper first identifies frequent Itemsets of Interest for Positive Association Rules and Infrequent Itemsets of Interest for Negative Association Rules using Algorithm All Itemsets of Interest based on minsupp, minconf and mininterset. And then, this paper extracts positive rules from frequent itemsets of interest and negative rules from infrequent itemsets of interest using Algorithm of Extracting Positive and Negative Association Rules. In Table 1, there are positive rules and negative rules for three databases: Grocery Stroe, Electronic and Book.

Table 1 Experimental Results

Database Name	Positive Rules	Negative Rules
Transaction (Grocery Store)	11	1
Transaction(Electronic)	15	3
Transaction(Book)	10	5

7. Efficiency

Table 2 shows the run time comparison of MBP (Mining By Pruning) and MNP (Mining with No-Pruning) in terms of efficiency in generating frequent itemsets. The MNP generate infrequent itemsets that satisfy conditions: (1) A∩B = Ø; (2) supp(A) ≥ minsupp; (3) supp(AU¬B) ≥ minsupp (or supp(¬AUB) ≥ minsupp, or supp(¬AU¬B) ≥ minsupp).So , The MNP does not have any specific pruning facility. The MBP is All Itemsets Of Interest procedure with the puring strategies in steps (3.4) and (3.5) in section (3.4) that remove all uninteresting itemsets that not satisfy interestingness constraints. Efficiency show the running time of MNP and MBP in miniseconds in generating frequent itemsets.

Table 2 Efficiency

Database Name	Min Support	MBP	MNP
Transaction (Grocery store)	0.3	232 ms	299 ms
Transaction (Electronic)	2	296 ms	326 ms
Transaction (Book)	1	245 ms	305 ms

8. Conclusion

Association rules are traditionally defined as implications of the form A ⇒ B□where A and B are frequent itemsets in a transaction database. This paper extends this definition to include association rules A ⇒¬□B,□□¬A ⇒ B □and ¬A ⇒□¬B□□which indicate negative associations between itemsets, The rules of the form A ⇒ B□ is positive rules and the rules of the other forms are negative rules.

This paper has constructed a new method for mining both positive and negative association rules in databases. Some infrequent itemsets are of interest in the method but not in existing research efforts. This paper has defined a set of conditions for frequent itemsets and infrequent itemsets to be of interest, and has used the increasing degree of the conditional probability relative to the prior probability to estimate the confidence of positive and negative association rules. Experimental results have demonstrated that proposed approach is efficient and promising.

10. References

[1] A. Savasere, E. Omiecinski, and S. Navathe, (1998) Mining for Strong Negative Associations in a large Database of Customer Transactions.

[2] C. Zhang and S. Zhang, “Mining both positive and Negative Association Rules”, University of Technology, Sydney.@IT.UTS.EDU.AU.

[3] J. Han and M. Kamber, “Data Mining Concepts and Technique”, ISBN 1-55860-489-8, Morgan Kaufmann Publishers.

[4] Pang-Ning Tan, Michale Steinbach.Vipi Kumar “Introduction to Data mining”.

[5] R. Agrawal and R. Srikant, “Fast Algorithm for Mining Association Rules in Large Database”. In processing of the 20th international Conference on Very Large Database, pp407-419, Santiago, Child, 1994.