Discovering Associations in Very Large Databases by Approximating*

Shichao Zhang[†] and Chengqi Zhang[‡]

Abstract

Mining association rules has posed great challenge to the research community. Despite efforts in designing fast and efficient mining algorithms, it remains a time consuming process for very large databases. In this paper, we adopt a slightly different approach to this problem, which can mine approximate association rules quickly. By considering the database as a set of records that are randomly appended, we can apply the central limit theorem to estimate the size of a random subset of the database, and discover both positive and negative association rules by generating all possible useful itemsets from the random subset. However, because of approximation errors, it is possible for some valid rules to be missed, while other invalid rules may be generated. To deal with this problem, we adopt a two phase approach. First, we discover all promising approximate rules from a random sample of the database. Second, these approximate results are used as heuristic information in an efficient algorithm that requires only one-pass of the database to validate rules that have support and confidence close to the desired support and confidence values. We evaluated the proposed technique, and our experimental results demonstrate that the approach is efficient and promising.

Keywords: Data mining, data processing, approximating rule, assisting knowledge discovery, data analysis.

1 Introduction

One of the main challenges in data mining is to identify association rules for very large databases that comprise millions of transactions and items. Some recent efforts have focused on designing efficient algorithms [2, 4, 7, 15], employing partitioning techniques [6, 9, 14], supporting incremental updating and exploiting parallelism [10, 13, 16]. The main "limitation" of these approaches, however, is that

*This research is partial supported by a large grant from the Australian Research Council (DP0343109) and partial supported by large grant from the Guangxi Natural Science Funds

[†]Faculty of Information Technology, University of Technology, Sydney, PO Box 123, Broadway NSW 2007, Australia, and Guangxi Teachers University, Gulin, P R China. Email: zhangsc@it.uts.edu.au

[‡]Faculty of Information Technology, University of Technology, Sydney, PO Box 123, Broadway NSW 2007, Australia. Email: chengqi@it.uts.edu.au

they require multiple passes over the database. For very large databases that are typically disk resident, this requires reading the database completely for each pass resulting in a large number of disk I/Os.

An alternative approach is to take a sample of the database, and determine association rules that are valid on the sample database. In other words, the problem of mining the association rules becomes a 3-step procedure:

(1) Generate a random subset of a given large database;

(2) Generate all large itemsets in the random subset;

(3) Generate all the rules with both support and confidence greater than or equal to minimum support and minimum confidence respectively.

As the sample size is typically very much smaller than the original database size, the association rules on the sample can be obtained at a much faster time. We shall refer to these association rules (obtained from the sample) as *approximate association rules*. The key issue in this approach is to pick a right sample that is representative of the database, so that the approximate association rules are indeed the association rules that hold on the database.

In this paper, we reexamine mechanisms for the 3 steps discussed above. To obtain a random sample of the database, we apply the central limit theorem. As we shall see shortly, the use of the central limit theorem allows us to cut down the sample size by about half compared to known techniques [11, 12]. For the second subtask, a new algorithm for generating all possible useful itemsets for mining rules with both positive and negative itemsets is proposed. Finally, the last subtask is solved by generating all positive and negative association rules.

Unfortunately, because of approximation errors, it is possible for some valid rules to be missed, while other invalid rules may be generated. To deal with this problem, we adopt a two phase approach. First, we discover all promising approximate rules from a random sample of the database. Second, these approximate results are used as heuristic information in an efficient algorithm that requires only one-pass of the database to validate rules that have support and confidence close to the desired support and confidence values.. We evaluated the proposed technique, and our experimental results demonstrate that the approach is efficient and promising.

The rest of this paper is organized as follows. In the next section, we briefly review some concepts and definitions. In Section 3, we apply the central limit theorem to mine approximate association rules. In order to discover both of positive and negative association rules, an algorithm to generate all possible useful itemsets is also proposed. In Section 4, we evaluate the effectiveness of the proposed approach experimentally. In Section 5, we propose a method to (1) assist knowledge discovery and (2) determine the validation of the rules with support or confidence close to the user-specified thresholds. Finally, we summarize our contributions in section 6.

2 Basic Concepts

One of the most widely used data mining model for association rules is the support-confidence framework established by Agrawal, Imielinski, and Swami [1]. We shall review some of the concepts here.

Let $I = \{i_1, i_2, \cdots, i_N\}$ be a set of N distinct literals called *items.* D is a set of variable length transactions over I. A transaction is a set of items, i.e., a subset of I. A transaction has an associated unique identifier called TID.

In general, a set of items (such as the antecedent or the consequent of a rule) is referred to as an *itemset.* For simplicity, an itemset $\{i_1, i_2, i_3\}$ is sometimes written as $i_1 i_2 i_3$.

For an itemset $A \subseteq I$ and a transaction $T \in D$, A is purchased (occurred) in T (or T contains A) if $\forall a \in A(\exists i((1 \leq i \leq n) \wedge (T(i) = a)))$, where '$T(i)$' is ith element of T.

The number of items in an itemset is the *length* (or the *size*) of an itemset. Itemsets of some length k are referred to as a k-itemsets.

An itemset has an associated measure of statistical significance called *support*, denoted as *supp*. For an itemset $A \subseteq I$, $supp(A) = s$, if the fraction of transactions in D containing A equals s. An itemset A is a large itemset if $supp(A) \geq min_{supp}$, where 'min_{supp}' is a user specified minimum support.

While A indicates the occurrence of an itemset A, the *negation* of A means the nonoccurrence of A, stood for $\overline{A}$ [1]. The support of $\overline{A}$ is as $supp(\overline{A}) = 1 - supp(A)$. Generally, for itemsets $A = \{i_1, \cdots, i_m\}$ and $B = \{j_1, \cdots, j_n\}$, the support of $\overline{A} \cup B$ is $supp(\overline{A} \cup B) = supp(B) - supp(A \cup B) = supp(\{j_1, \cdots, j_n\}) - supp(\{i_1, \cdots, i_m, j_1, \cdots, j_n\})$.

An *association rule* is an implication of the form $A \Rightarrow B$ (or written as $A \rightarrow B$), where $A, B \subseteq I$, and $A \cap B = \emptyset$. A is called the *antecedent* of the rule, and B is called the *consequent* of the rule.

An association rule $A \rightarrow B$ has a measure of its strength called *confidence* (denoted as $conf$) defined as the ratio $supp(A \cup B)/supp(A)$, where $A \cup B$ means that both A and B are present in transactions.

The work in this paper extends traditional associations to include association rules of forms $A \rightarrow \overline{B}$, $\overline{A} \rightarrow B$, and $\overline{A} \rightarrow \overline{B}$, which indicate negative associations between itemsets. We call rules of the form $A \rightarrow B$ *positive association rules*, and rules of the other forms *negative association rules.* Negative rules indicate that the presence of some itemsets will imply the absence of other itemsets in the same transactions. Negative rules are also very useful in association analysis, although they are hidden and different from positive association rules.

The problem of mining association rules is to generate all rules $A \rightarrow B$ that have both support and confidence greater than or equal to some user specified minimum support (min_{supp}) and minimum confidence (min_{conf}) thresholds respectively, i.e.

[1] An itemset A is often taken as an event in computations meaning that A is true in a transaction if item i presents in the transaction for $\forall i \in A$. $\overline{A}$ is taken as an event in computations meaning that $\overline{A}$ is true in a transaction if item i does not present in the transaction for $\exists i \in A$. That is, $\overline{A}$ is different from $I - A$.

for regular associations:

$$supp(A \cup B) \geq min_{supp}, \qquad conf(A \to B) = \frac{supp(A \cup B)}{supp(A)} \geq min_{conf}.$$

It can be decomposed into the following two subproblems.

(1) All itemsets that have support greater than or equal to the user specified minimum support are generated. That is, generating all large itemsets.

(2) Generate all the rules that have minimum confidence in the following naive way: For every large itemset X and any $B \subset X$, let $A = X - B$. If the rule $A \to B$ has the minimum confidence (or $supp(X)/supp(A) \geq min_{conf}$), then it is a valid rule.

Example 1. *Let $T_1 = \{i_1, i_2, i_4\}$, $T_2 = \{i_1, i_2, i_4\}$, $T_3 = \{i_2, i_3, i_4\}$, $T_4 = \{i_2, i_3, i_4\}$, and $T_5 = \{i_1, i_2\}$ be the only transactions in a database. Let the minimum support and minimum confidence be 0.6 and 0.85 respectively. Then the large itemsets are the following: $\{i_1\}$, $\{i_2\}$, $\{i_4\}$, $\{i_1, i_2\}$ and $\{i_2, i_4\}$. The valid rules are $i_1 \to i_2$ and $i_4 \to i_2$.*

3 Mining Approximate Rules

In probability theory, if a situation is such that only two outcomes, often called success and failure, are possible, it is usually called a *trial*. The variable element in a trial is described by a probability distribution on a sample space of two elements, 0 representing failure and 1 success; this distribution assigning the probability $1-\theta$ to 0 and θ to 1, where $0 \leq \theta \leq 1$. Suppose we consider n independent repetitions of a given trial. The variable element in these is described by a probability distribution on a sample space of 2^n points, the typical point being $x = (x_1, x_2, \cdots, x_n)$, where each x_i is 0 or 1, and x_i represents the result of the ith trial. The appropriate probability distribution is defined by

$$p_\theta(x) = \theta^{m(x)}(1-\theta)^{n-m(x)},$$

where $m(x) = \sum_{i=1}^{n} x_i$ is the number of 1s in the results of the n trials, this being so since the trials are independent.

Given an x in this situation it seems reasonable to estimate θ by $m(x)/n$, the proportion of successes obtained. This seems in some sense to be a 'good' estimate of θ.

In data mining, a database D can be taken as a trial. For any itemset A, it is 1 if the itemset A occurs in a transaction T (written as $T(A)$), else it is 0 (written as $\neg T(A)$). Let P be the set of all transactions that the itemset A occurs in, and Q be the set of all transactions that the itemset A doesn't occur in. Then P and Q are partitions of D as follows.

$$P = \{T | T(A)\},$$
$$Q = \{T | \neg T(A)\}.$$

Given a database, its n transactions can be viewed as n independent data stored in the database. Certainly, each transaction has two possible outcomes for an itemset A, which are 1 and 0. Suppose the probability of A occurring in the database is p and the probability of A not occurring is $q = 1 - p$. Since the database is static, we can say that probability p of A occurring in each transaction is the same for each transaction. Hence, this given database can be taken as a Bernoulli trial.

3.1 The Application of Central Limit Theorem

The central limit theorem is one of the most remarkable results in probability theory. Loosely put, it states that the sum of a large number of independent random variables has a distribution that is approximately normal. Hence it not only provides a simple method for computing approximating probabilities for sums of independent random variables, but it also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit bell-shaped (that is, normal) curves. In its simplest form the central limit theorem is as follows.

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent and identically distributed random variables, each having finite mean $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Then the distribution of

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is,

$$P\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx \quad as \ n \to \infty. \tag{1}$$

Readers are referred to [5] for other concepts and theorems.

We now set up a new mining model in this subsection, which applies central limit theorem to mine approximate association rules from large databases.

Theorem 1. *Let I be the set of items in database D, $A \subseteq I$ an itemset, $\eta > 0$ the degree of asymptotic to association rules and $\xi \geq 0$ the upper probability of $P[|Ave(X_n) - p| \leq \eta]$, where $Ave(X_n)$ is the average of A occurring in n transactions in D and p is the probability of A in D. Suppose records in D are matched Bernoulli trials. If n random records of D is enough for determining the approximate association rules in D according to central limit theorem, n must be as follows:*

$$n \geq \frac{z^2((1+\xi)/2)}{4\eta^2} \tag{2}$$

where $z(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$ is a standard normal distribution function, which can find out it from the Appendix in [5].

Proof. From the given conditions in this theorem, we take

$$P(|Ave(X_n) - p| \leq \eta) = \xi$$

Clearly,

$$\begin{aligned} P(|Ave(X_n) - p| \leq \eta) &= P(-\eta \leq (Ave(X_n) - p) \leq \eta) \\ &= P(\frac{-\eta}{1/(2\sqrt{n})} \leq \frac{Ave(X_n) - p}{1/(2\sqrt{n})} \leq \frac{\eta}{1/(2\sqrt{n})}) \\ &\approx N(2\eta\sqrt{n}) - N(-2\eta\sqrt{n}) \\ &= 2N(2\eta\sqrt{n}) - 1 \end{aligned}$$

where $N()$ is the distribution function of the standard normal distribution. And for this probability to equal ξ we need

$$N(2\eta\sqrt{n}) = \frac{1}{2}(1+\xi)$$

which is satisfied by

$$2\eta\sqrt{n} = z((1+\xi)/2)$$

the required value for n then is

$$n \geq \frac{z^2((1+\xi)/2)}{4\eta^2}$$

□

Example 2. *Suppose a new process is available for doping silicon chips, used in electronic devices. p (unknown) is the probability that each chip produced in this way is defective. We assume that the defective chips are independent of each other. How many chips, n, must we produce and test so that the proportion of defective chips found ($Ave(X_n)$) does not differ from p by more than* 0.01, *with probability at least* 0.99*? That is, we want n such that*

$$P(|Ave(X_n) - p| < 0.01) > 0.99,$$

$\eta = 0.01, \xi = 0.99, z(0.995) = 2.57$, *we have*

$$n = \frac{2.57^2}{4 * 0.01^2} = 16513,$$

considerably smaller than the value n = 27000 *that is needed by using the model in Chernoff bounds [11, 12].*

Based on Theorem 1, the random target database can be obtained in two steps: (1) generating a set X of pseudo-random numbers, where $|X| = n$ and (2) generating the random database RD (instance set) from D using pseudo-random number set X. That is, for any $x_i \in X$, get $(x_i + 1)^{th}$ record of D and append it into RD.

Note that generating random database RD of the given database D doesn't mean to establish a new database RD. It only needs to build a view RD over D.

3.2 Mining Approximate Association Rules

In this subsection, we construct a new model for discovering both of positive and negative association rules. For this goal, an algorithm of generating all positive and negative large itemsets is also proposed.

Positive and Negative Large Itemsets

For mining general approximate association rules, all positive and negative large itemsets in a random database would be generated. For example, if one of $A \rightarrow \overline{B}$, $\overline{A} \rightarrow B$ and $\overline{A} \rightarrow \overline{B}$ can be discovered, then one of $supp(A \cup \overline{B}) \geq min_{supp}$, $supp(\overline{A} \cup B) \geq min_{supp}$ and $supp(\overline{A} \cup \overline{B}) \geq min_{supp}$ must hold. This means that $supp(A \cup B) < min_{supp}$. However, itemsets such as $A \cup B$, are not generated as large itemsets into the set of all large itemsets. In order to mine negative rules, we present a procedure to generate all positive and negative large itemsets in a random database as follows.

Procedure 1. *PNLargeItemsets*

```
  Input: D: database; $min_{supp}$: minimum support;
  Output: PL: large itemsets; NL: negative large itemsets;
  Begin
 (1) generate sample RD of a given database D;
     let $PL \leftarrow \emptyset$; $NL \leftarrow \emptyset$;
 (2) let $L_1 \leftarrow \{$large 1-itemsets$\}$; $PL \leftarrow PL \cup L_1$;
 (3) for $(k = 2;\ (L_{k-1} \neq \emptyset);\ k++)$ do
     begin //Generate all possible positive and negative k-itemsets of interest in RD.
   (3.1)  let $L_k \leftarrow \{\{x_1, \ldots, x_{k-2}, x_{k-1}, x_k\} \mid \{x_1, \ldots, x_{k-2}, x_{k-1}\} \in L_{k-1} \wedge \{x_1, \ldots x_{k-2}, x_k\} \in L_{k-1}\}$;
   (3.2)  for each transaction t in RD do
             begin
          //Check which k-itemsets are included in transaction t.
              let $t_{Tem} \leftarrow$ the k-itemsets in t that are also contained in $L_k$;
              for each itemset A in $t_{Tem}$ do
                let $A.count \leftarrow A.count + 1$;
             end
   (3.3)  //Selecting all positive k-itemsets in $L_k$
          let $Tem_k \leftarrow \{C | C \in L_k \wedge (supp(C) = (C.count/|RD|) >= min_{supp})\}$;
          let $PL \leftarrow PL \cup Tem_k$;
          //Selecting all negative k-itemsets in $L_k$
            let $NL \leftarrow NL \cup (L_k - Tem_k)$;
     end
 (4) output PL and NL;
  End.
```

The procedure *PNLargeItemsets* generates all positive and negative itemsets in the sample *RD*. The initialization and generating sample *RD* of a given database *D* are done in Step (1). Step (2) counts the frequencies of itemsets in *RD*. Step (3) generates all positive and negative itemsets of interest.

Rules of Interest

In [8], Piatetsky-Shapiro argued that a rule $X \rightarrow Y$ is not interesting if

$$supp(X \rightarrow Y) \approx supp(X)supp(Y)$$

According to probability interpretation [3]: $supp(X \cup Y) = P(X \cup Y)$ and $conf(X \rightarrow Y) = P(Y|X) = P(X \cup Y)/P(X)$ Then Piatetsky-Shapiro's argument can be denoted as

$$P(X \cup Y) \approx P(X)P(Y).$$

This means that $X \rightarrow Y$ cannot be extracted as a rule if $P(X \cup Y) \approx P(X)P(Y)$. Actually, $P(X \cup Y) \approx P(X)P(Y)$ denotes X is approximately independent to Y in probability theory. A statistical definition [3] of dependence of the sets X and Y is

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

with the obvious extension to more than two sets. This formula is referred to as the *interest* of Y given X is one of the main measurements of uncertainty of association rules. Certainly, the further the value is from 1, the more the dependence. Or for $1 > \lambda > 0$, if $|\frac{P(X \cup Y)}{P(X)P(Y)} - 1| \geq \lambda$, then $X \rightarrow Y$ is a rule of interest.

By Piatetsky-Shapiro's argument, we can divide $Interest(X, Y)$ into 3 cases as follows:

(1) if $P(X \cup Y)/(P(X)P(Y)) = 1$, then $P(X \cup Y) = P(X)P(Y)$ or Y and X are independent;

(2) if $P(X \cup Y)/(P(X)P(Y)) > 1$, or $P(X \cup Y) > P(X)P(Y)$, then Y is positively dependent to X;

(3) if $P(X \cup Y)/(P(X)P(Y)) < 1$, or $P(X \cup Y) < P(X)P(Y)$, then Y is negatively dependent to X, or $\overline{Y}$ is positively dependent to X.

In this way, we can define another form of interpretation of rules of interest as follows. For $1 > \lambda > 0$, (a) if $\frac{P(X \cup Y)}{P(X)P(Y)} - 1 \geq \lambda$, then $X \rightarrow Y$ is a rule of interest; and (b) if $-(\frac{P(X \cup Y)}{P(X)P(Y)} - 1) \geq \lambda$, then $X \rightarrow \overline{Y}$ is a rule of interest.

Theorem 2. *Let I be the set of items in database D, $X, Y \subseteq I$ be itemsets, $X \cap Y = \emptyset$, $P(X) \neq 0$ and $P(Y) \neq 0$. min_{supp}, min_{conf} and $\lambda > 0$ are given by users or experts. If*

(1) $supp(X \cup Y) \geq min_{supp}$, $conf(X \rightarrow Y) \geq min_{conf}$, and $P(X \cup Y) - P(X)P(Y) \geq \lambda$, then $X \rightarrow Y$ can be extracted as a rule of interest.

(2) $supp(X \cup \overline{Y}) \geq min_{supp}$, $supp(Y) \geq min_{supp}$, $conf(X \rightarrow \overline{Y}) \geq min_{conf}$, and $-(P(X \cup Y) - P(X)P(Y)) \geq \lambda$, then $X \rightarrow \overline{Y}$ can be extracted as a rule of interest.

Proof. From assumption of the above theorem, we have

$$\frac{|(P(X \cup Y) - P(X)P(Y))|}{P(X)P(Y)} \geq \frac{\lambda}{P(X)P(Y)},$$

or

$$|\frac{P(X \cup Y)}{P(X)P(Y)} - 1| \geq \frac{\lambda}{P(X)P(Y)}.$$

Because $0 < P(X)P(Y) \leq 1$, so $\lambda/(P(X)P(Y)) \geq \lambda$. Hence,

$$|\frac{P(X \cup Y)}{P(X)P(Y)} - 1| \geq \lambda,$$

That is, $X \rightarrow Y$ can be extracted as a rule of interest. □

Mining Positive And Negative Rules

By our definition on interest, if $P(X \cup Y) \approx P(X)P(Y)$, X is approximately independent to Y in probability theory; if the greater the value of $P(X \cup Y) - P(X)P(Y) > 0$ is, the more the positive dependence; and if the smaller the value of $P(X \cup Y) - P(X)P(Y) < 0$ is, the more the negative dependence. However, $-P(X)P(Y) \leq P(X \cup Y) - P(X)P(Y) \leq P(X)(1 - P(Y))$. In order to reflect this relationship between $P(X \cup Y)$ and $P(X)P(Y)$, we propose the *probability ratio* (PR) model here. Under the PR model, we define the measure PR to determine the degree in which the valid rule $X \rightarrow Y$ is interesting.

$$PR(Y|X) = \begin{cases} \frac{P(X \cup Y) - P(X)P(Y)}{P(X)(1 - P(Y))}, & if \ \ P(X \cup Y) \geq P(X)P(Y), \\ & \quad\ \ P(X)(1 - P(Y)) \neq 0. \\ \frac{P(X \cup Y) - P(X)P(Y)}{P(X)P(Y)}, & if \ \ P(X \cup Y) < P(X)P(Y), P(X)P(Y) \neq 0. \end{cases}$$

Certainly, PR has some properties as follows.

Property 1. *PR satisfies the following:*

$$PR(Y|X) + PR(\overline{Y}|X) = 0.$$

Proof. We shall only prove the property holds when $P(X \cup Y) \geq P(X)P(Y)$. The others can be derived in a similar manner. Since

$$P(X \cup Y)/P(X) = P(Y|X), P(X \cup \overline{Y})/P(X) = P(\overline{Y}|X), P(Y|X) + P(\overline{Y}|X) = 1,$$

and

$$P(\overline{Y}|X) = 1 - P(Y|X) \leq 1 - P(Y) = P(\overline{Y}).$$

Therefore,

$$PR(Y|X) = \frac{P(X \cup Y) - P(X)P(Y)}{P(X)(1 - P(Y))} = \frac{P(Y|X) - P(Y)}{1 - P(Y)},$$

$$PR(\overline{Y}|X) = \frac{P(X \cup \overline{Y}) - P(X)P(\overline{Y})}{P(X)P(\overline{Y})} = \frac{P(\overline{Y}|X) - P(\overline{Y})}{P(\overline{Y})}.$$

Hence,

$$PR(Y|X) + PR(\overline{Y}|X) = \frac{P(X \cup Y) - P(X)P(Y)}{P(X)(1 - P(Y))} + \frac{P(X \cup \overline{Y}) - P(X)P(\overline{Y})}{P(X)P(\overline{Y})}$$

$$= \frac{P(Y|X) - P(Y)}{1 - P(Y)} + \frac{P(\overline{Y}|X) - P(\overline{Y})}{P(\overline{Y})}$$

$$= \frac{P(Y|X) - P(Y)}{1 - P(Y)} + \frac{(1 - P(Y|X)) - (1 - P(Y))}{1 - P(Y)} = 0.$$

So, we have $PR(Y|X) + PR(\overline{Y}|X) = 0$. □

We now apply the PR model to measure the uncertainties of association rules.

(1) For an association rule $A \rightarrow B$, its *supp*$(A \cup B)$ is $P(A \cup B)$ and, $PR(B|A)$ is taken as the confidence of the rule. The task of mining this association rule is defined as follows. For itemset $A \cup B$, if *supp*$(A \cup B) \geq min_{supp}$ and $PR(B|A) \geq min_{conf}$, then $A \rightarrow B$ can be extracted as a valid rule.

(2) For an association rule $A \rightarrow \overline{B}$, $PR(\overline{B}|A) = -PR(B|A)$ according to Property 1. Therefore, if *supp*$(A \cup \overline{B}) \geq min_{supp}$, *supp*$(B) \geq min_{supp}$, $PR(B|A) < 0$ and $PR(\overline{B}|A) \geq min_{conf}$, then $A \rightarrow \overline{B}$ can extracted as a valid rule.

(3) For $\overline{A} \rightarrow B$, $PR(B|\overline{A})$ is taken as the confidence of the rule. The task of mining this association rule is defined as follows. For itemset $\overline{A} \cup B$, if *supp*$(A) \geq min_{supp}$, *supp*$(\overline{A} \cup B) \geq min_{supp}$ and $PR(B|\overline{A}) \geq min_{conf}$, then $\overline{A} \rightarrow B$ can be extracted as a valid rule.

(4) For an association rule $\overline{A} \rightarrow \overline{B}$, $PR(\overline{B}|\overline{A}) = -PR(B|\overline{A})$ according to Property 1. Therefore, if *supp*$(\overline{A} \cup \overline{B}) \geq min_{supp}$, *supp*$(A) \geq min_{supp}$, *supp*$(B) \geq min_{supp}$, $PR(B|\overline{A}) < 0$ and $PR(\overline{B}|\overline{A}) \geq min_{conf}$, then $\overline{A} \rightarrow \overline{B}$ can extracted as a valid rule.

Note that the requirements that $supp(B) \geq min_{supp}$ and $supp(A) \geq min_{supp}$ ensure the probability significance of rules with negative itemsets.

We now demonstrate how to apply this model to discover association rules with the data in Example 1. Let $min_{supp} = 0.2$ and $min_{conf} = 0.4$.

Example 3. *For itemset $A \cup C$, $P(A) = 0.6$, $P(C) = 0.4$ and $P(A \cup C) = 0$, we have $P(A \cup C) < P(A)P(C)$. This means that the disbelief increases, or $A \rightarrow \overline{C}$ may be extracted as a rule of interest. Furthermore,*

$$PR(C|A) = \frac{P(A \cup C) - P(A)P(C)}{P(A)P(C)} = \frac{0 - 0.6 \bullet 0.4}{0.6 \bullet 0.4} = -1,$$

According to our model, $A \rightarrow \overline{C}$ can be extracted as a valid rule due to $PR(\overline{C}|A) = -PR(C|A) = 1 > min_{conf}$, $supp(A \cup \overline{C}) = 0.6 > min_{supp}$, and $supp(C) = 0.4 > min_{supp}$.

As we have seen, our PR model is both reasonable and comprehensive. And general association rules can be easily discovered. Furthermore, we can obtain the following theorem that facilitates the extraction of interesting rules.

Theorem 3. *Let I be the set of items in database D, $X, Y \subseteq I$ be itemsets, $X \cap Y = \emptyset$, $P(X) \neq 0$ and $P(Y) \neq 0$. min_{supp}, min_{conf} and $\lambda > 0$ are given by users or experts. Then*

(1) if $supp(X \cup Y) \geq min_{supp}$ and $PR(Y|X) \geq Max\{min_{conf}, \lambda\}$, then $X \rightarrow Y$ can be extracted as a rule of interest;

(2) if $supp(X \cup \overline{Y}) \geq min_{supp}$, $supp(Y) \geq min_{supp}$ and $PR(\overline{Y}|X) \geq Max\{min_{conf}, \lambda\}$, then $X \rightarrow \overline{Y}$ can be extracted as a rule of interest;

(3) if $supp(\overline{X} \cup Y) \geq min_{supp}$, $supp(X) \geq min_{supp}$ and $PR(Y|\overline{X}) \geq Max\{min_{conf}, \lambda\}$, then $\overline{X} \rightarrow Y$ can be extracted as a rule of interest;

(4) if $supp(\overline{X} \cup \overline{Y}) \geq min_{supp}$, $supp(Y) \geq min_{supp}$, $supp(X) \geq min_{supp}$ and $PR(\overline{Y}|\overline{X}) \geq Max\{min_{conf}, \lambda\}$, then $\overline{X} \rightarrow \overline{Y}$ can be extracted as a rule of interest.

Proof. As before, we only prove (1) of the above theorem since the rest can be obtained similarly. We first prove that (1) holds. Since $PR(Y|X) \geq Max\{min_{conf}, \lambda\}$, according to the assumption in (1), we have $PR(Y|X) \geq min_{conf}$ and $PR(Y|X) \geq \lambda$.

On the other hand, because $PR(Y|X) \geq 0$, and using the Property 1, $PR(Y|X) + PR(\overline{Y}|X) = 0$, we have $-PR(\overline{Y}|X) \geq \lambda$.

According to previous interpretation of rules of interest, $X \rightarrow \overline{\overline{Y}}$ can be extracted as a rule of interest. That is

$$X \rightarrow Y$$

can be extracted as a rule of interest. □

Algorithm

Let D be a database, $|D|$ the total number of transactions in D, I the set of all items in D, and for $X \subseteq I$, $|X|$ the number of transactions in D that contain itemset X, min_{supp}, min_{conf}, λ and γ given by users. The algorithm of discovering association rules in our probability ratio model is constructed as follows.

Algorithm 1. *PRModel*

```
Input: D: database, $min_{supp}, min_{conf}$, $\lambda$ and $\gamma$: threshold values;
Output: approximate rules;
(1) Determine the sample size, n, based on the central limit theorem;
    Generate the sample database with n transactions;
    call routine PNLargeItemsets;
(2) for any large itemset $A$ in $PL$ begin
        for any itemset $X \subset A$ begin
          let $Y = A - X$;
          if $|PR(Y|X)| \geq Max\{min_{conf}, \lambda\}$ then
            output the rule $X \to Y$
              with confidence $PR(Y|X)$ and support $P(A)$;
          end
    for any itemset $A$ in $NL$ begin
        for any itemset $X \subset A$ begin
          let $Y = A - X$;
          if ($supp(X \cup \overline{Y}) \geq min_{supp}$ and $supp(Y) \geq min_{supp}$
              and $|PR(\overline{Y}|X)| \geq Max\{min_{conf}, \lambda\}$) then
            output the rule $X \to \overline{Y}$
              with confidence $PR(\overline{Y}|X)$ and support $P(A)$;
          end
          if ($supp(\overline{X} \cup Y) \geq min_{supp}$ and $supp(X) \geq min_{supp}$
              and $|PR(Y|\overline{X})| \geq Max\{min_{conf}, \lambda\}$) then
            output the rule $\overline{X} \to Y$
              with confidence $PR(Y|\overline{X})$ and support $P(A)$;
          end
          if ($supp(\overline{X} \cup \overline{Y}) \geq min_{supp}$ and $supp(X) \geq min_{supp}$
              and $supp(Y) \geq min_{supp}$ and $|PR(\overline{Y}|X)| \geq Max\{min_{conf}, \lambda\}$)
            then
            output the rule $\overline{X} \to \overline{Y}$
              with confidence $PR(\overline{Y}|\overline{X})$ and support $P(A)$;
          end
        end
    end
    endall.
```

Algorithm *PRModel* generates all positive association rules in PL and negative association rules in NL. Step (1) calls procedure *PNLargeItemsets* to generate the sets PL and NL with positive and negative large itemsets in the database D. Step (2) firstly generates positive association rules of interest of the form: $X \to Y$, in PL. If $PR(Y|X) \geq min_{conf}$, $X \to Y$ is extracted as a valid rule. If $PR(X|Y) \geq min_{conf}$, $Y \Rightarrow X$ is extracted as a valid rule. Secondly, the step generates negative association rules of interest of the forms $\overline{X} \to Y$, $Y \to \overline{X}$, $\overline{X} \to \overline{Y}$, and $\overline{Y} \to \overline{X}$, in NL.

4 An Experimental Study

To study the effectiveness of our model, we have performed several experiments. Our server is Oracle 8.0.3, and the algorithm is implemented on Sun SparcServer using Java, and JDBC API is used as the interface between the program and Oracle. The database used in our experiments has the following conceptual scheme

$$Report(sno, test, grade, area)$$

where *sno* is the primary key about student numbers, *test* is an attribute about examinations of subjects, *grade* is an attribute about students' grades with (A, B, C, D, E) as its domain, *area* is an attribute about students' nationality with a domain $(China, Singapore, \cdots)$. In order to illustrate the efficiency of our approximate rule model, we list partially the experimental results, which are the large itemsets and their supports. For more details, please refer to Appendix A.

Let $min_{supp} = 0.2$ and $min_{conf} = 0.6$. Some results are listed in Table 1.

We evaluated three methods: the traditional approach where the entire database is used (denoted D), the sampling approach based on Chernoff bounds [11, 12] (denoted LRD), and the proposed approach using the central limit theorem (denoted CRD). As shown in Table 1, the supports for the various useful itemsets are very close among the three methods. For example, the supports of item "China" are 37%, 36.78% and 36.48% for D, LRD and CRD respectively. This shows that relevant itemsets can be determined based on a small sample of the database. In our case, LRD requires only 15000 records which is only 15% of the original database size, while CRD makes use of no more than 7% of the original database size. We also note that the running time of mining the original database is 815 seconds. The time for LRD is 436 seconds (consisting of 207 seconds for LRD and 229 seconds for approximate rules), while that of CRD is only 241 seconds (consisting 101 seconds for LRD and 140 seconds for approximate rules). The significant reduction is clearly due to the smaller size of the samples. We also note that CRD is more efficient than LRD, making CRD a promising approach for mining association rules.

Referring to the Table, some of the rules of interest are $China \to B$, $China \to \overline{C}$, $China \to \overline{Singapore}$, $Singapore \to C$, $B \to \overline{C}$. However, from the example, we also note the following problems, which we shall investigate shortly.

(i) Some rules such as $China \to \overline{Singapore}$ and $B \to \overline{C}$ are also extracted as rules of interest.

(ii) Due to the probability significance and the constraint condition of min_{supp}, some rules such as $China \rightarrow \overline{D}$, $Singapore \rightarrow \overline{D}$, $China \rightarrow \overline{E}$ and $Singapore \rightarrow \overline{E}$, can't be extracted as negative rules of interest in our model. In some context, these rules are useful for applications. But mining rules such as $China \rightarrow \overline{Tom}$ has no significance, where "Tom" is name of some student.

Table 1: Some itemsets in the original database.

DB	useful Itemset	Support	size of sample	running time
D	China	37%	100000	815
	Singapore	50%		
	B	33.2%		
	C	42.05%		
	China, B	27.75%		
	Singapore, C	35%		
	China, Singapore	0%		
	China, C	3.1%		
	B, C	0%		
LRD	China	36.78%	15000	436
	Singapore	50.43%		
	B	33.43%		
	C	42.17%		
	China, B	27.83%		
	Singapore, C	34.97%		
	China, Singapore	0%		
	China, C	2.87%		
	B, C	0%		
CRD	China	36.48%	6724	241
	Singapore	50.82%		
	B	33, 45%		
	C	42.3%		
	China,B	27.71%		
	Singapore, C	35.07%		
	China, Singapore	0%		
	China, C	3.01%		
	B, C	0%		

As we have seen, if all data are randomly appended in to a given large database, the association rules can be approximated by our model using central limit theorem. The experiments also show the effectiveness of the proposed approach. Before closing this section, we shall make the following claim.

Claim 1. *Consider the given database D, we have*

(1) If all data are randomly appended into a given large database, association rules can be approximated by our model using limit theorems.

(2) If $A \rightarrow B$ can extracted as a rule in our model, it must be a rule of interest.

(3) The model in central limit theorem is more efficient than the model in Chernoff bounds.

We now explain these arguments. (1) can directly be proven by the above algorithm and Theorem 1; (2) can be obtained from Theorem 3 and Algorithm 1.

For Theorem 1 and model based on Chernoff bounds [11, 12], we can compare the efficiency between Chernoff bounds and central limit theorem as follows.

$$\frac{1}{2\eta^2} ln \frac{2}{1-\xi} || \frac{z^2_{(1+\xi)/2}}{4\eta^2},$$

where "||" is a comparison symbol, or

$$ln \frac{2}{1-\xi} || \frac{z^2_{(1+\xi)/2}}{2},$$

where $(1+\xi)/2 \geq 0.5$. According to the list of standard normal distribution function, the following inequality holds for $1 \geq \xi > 0$

$$ln \frac{2}{1-\xi} \geq \frac{z^2_{(1+\xi)/2}}{2}.$$

Hence,

$$\frac{1}{2\eta^2} ln \frac{2}{1-\xi} \geq \frac{z^2_{(1+\xi)/2}}{4\eta^2},$$

Thus, the model in central limit theorem is more efficient than the model in large number law, i.e., (3) in Claim holds.

□

5 Assisting Knowledge Discovery

As has been shown, our model is efficient to discover approximate association rules in large databases. However, if the support of an itemset A is in the neighbour of min_{supp}, then A can be sometimes be treated as a large itemset and sometimes not as a large itemset due to approximation errors. In other words, some such itemsets are large itemsets in D but not in RD, and some such itemsets are not large itemsets in D but they are large itemsets in RD. This is a weakness of our model. For example, consider a random subset RD of a given large database D. Let $min_{supp} = 0.2$ and the probability of error to be tolerated be 0.05. Let two itemsets A and B in D with probabilities (supports) 0.18 and 0.23 respectively. Assume also that A and B are generated with probabilities 0.21 and 0.194 respectively, in the random database RD. This means that A is a large itemset in RD and B is not a large itemset in RD due to approximating error 0.05. They are unexpected results.

On the other hand, if we cannot compromise the validity of mined rules, or when certain support and confidence are necessary for some applications, $\eta > 0$ can be expected to be much smaller. This implies that we have to end up with a very large sample of the database, which diminishes the gains of sampling.

However, because of the randomness of data in a given database, we can roughly generate a possible large itemset set at first. Then this set is used as heuristic information to obtain large itemsets with only one pass through the given database. In this way, we can use such heuristic information to (1) assist knowledge discovery where accuracy is important or certain support and confidence is desirable, and (2) determine if an itemset in the neighbour of min_{supp} in the random subset of a given database is a large itemset.

Definition 1. *If an itemset A in RD is greater than or equal to $min_{supp} - \eta$, then it is reasonable in probability to conjecture that A is a large itemset in the database D. And itemset such as A is called hopeful large itemset in D. Reversedly, if an itemset A in RD is less than $min_{supp} - \eta$, then it is reasonable and comprehensive in probability to believe that A is impossible as a large itemset in the database D.*

Apparently, assessing hopeful large itemset are not only useful to the itemsets in the neighbour of min_{supp}, but also efficient to assist non-approximate knowledge discovery in databases. We now present the algorithm of accomplishing such two tasks as follows.

Procedure 2. *TLargeItemset*

```
Input: $\eta$: accuracy of results, $\xi$: probability of requirements, $min_{supp}$: minimum support,
        D: original database, HLIsSet: set of hopeful large itemsets;
Output: LI: large itemsets D;
  Begin
   let $LI \leftarrow \emptyset$;
   for each transaction $\tau$ of D do
     for each itemset $\alpha$ of HLIsSet do
       if $\alpha \in \tau$ then
         let $Count_\alpha \leftarrow Count_\alpha + 1$;
   for each itemset $\alpha$ of HLIsSet do
     if $supp(\alpha) \geq min_{supp}$ then
       let $LI \leftarrow LI \cup \{\alpha\}$;
   output the set LI of all large itemsets in D;
  end;
```

Again, if the confidence of a rule $A \rightarrow B$ is in the neighbour of min_{conf}, then $A \rightarrow B$ can be sometimes extracted as a valid rule and sometimes not as a valid rule due to the approximate error. The problem of the neighbour of min_{conf} can be addressed using a similar method as that for the neighbour of min_{supp}.

Now, we can describe the model of applying our method to assist non-approximate knowledge discovery in databases as follows. For a given large database D, with the users specified min_{supp} and min_{conf}, the following steps are performed.

(1) Generate a random subset RD of D according to our model in this paper;

(2) Generate the set $HLIsSet$ of all hopeful large itemsets in RD with support greater than or equal to $max\{0, min_{supp} - approximate\ error\}$;

(3) Generate all large itemsets in D with support greater than or equal to min_{supp} according to the set of hopeful large itemsets and Procedure 2;

(4) Generate all the rules with both support and confidence greater than or equal to minimum support and minimum confidence respectively, according to the large itemsets in the given database.

Certainly, applying approximate results to assist knowledge discovery needs only rough estimation, such as $\eta = 0.01$ and $\xi = 0.9$ are enough to generate all hopeful large itemsets. On the other hand, Algorithm 1 is linear. It can be guaranteed by the following theorem.

Theorem 4. *For given large database D, let $n = |D|$, m be the time of generating the set $HLIsSet$ of all hopeful large itemsets in random subset RD of D. Then the time of generating all large itemset in D is at least $O(m + n^2)$.*

Proof. According to the above definition, Algorithm 1 and Procedure 2, it needs only one pass to read the given database D. And each reading takes $t + t'$ to read a transaction from D and count itemsets in $HLIsSet$, where t the time to read a transaction from D, and t' the time to count all itemsets in $HLIsSet$. So, n reading incurs time of $n(t + t')$. Hence, the time to generate all large itemsets in D is $m + n(t + t')$. Let t'' be the time to count an itemset. Then $t' = t''|HLIsSet|$ for general databases. Because $|HLIsSet|$ is at least $O(n)$, and t and t'' are two small constants, so $m + n(t + t') = m + n(t + t''|HLIsSet|)$ is at least $O(m + n^2)$. □

In order to handle the problem caused by both the neighbour of min_{supp} and the neighbour of min_{conf}, we can use two methods as follows. One of them is to take $max\{0, min_{supp} - \eta\}$ and $max\{0, min_{conf} - \eta\}$ as the minimum support and minimum confidence respectively, for applications that need only approximate results. If an application requires more accurate results or certain support and confidence, the following method can be performed.

For a given large database D, min_{supp} and min_{conf} are given by users.

(1) Generate a random subset RD of D;

(2) Generate all hopeful large itemsets in RD with support greater than or equal to $max\{0, min_{supp} - approximate\ error\}$;

(3) Generate the set $RSET$ of all the rules with both support and confidence greater than or equal to minimum support ($max\{0, min_{supp} - \eta\}$) and minimum confidence ($max\{0, min_{conf} - \eta\}$) respectively, according to the hopeful large itemsets in RD;

(4) For the subset PS of $RSET$ with both support and confidence in the neighbour of min_{supp} and the neighbour of min_{conf} respectively, generate the set VRS of all rules in PS that is valid in D;

(5) Output $(RSET - PS) \cup VRS$.

Theorem 5. *For given large database D, min_{supp} and min_{conf} are given by users. $A \rightarrow B$ can be extracted as an approximate rule in the above model if and only if $A \rightarrow B$ is a valid rule in D.*

Proof. We first prove ($\rightarrow$). According to the above assumption, if

(1) $supp(A \cup B) \geq max\{0, min_{supp} - \eta\}$; and

(2) $conf(A \rightarrow B) \geq max\{0, min_{conf} - \eta\}$;

hold in random subset RD of D. By (4) and (5) in the above definition, we can obtain

(i) $supp(A \cup B) \geq min_{supp}$; and

(ii) $conf(A \rightarrow B) \geq min_{conf}$;

This means, $A \rightarrow B$ is still a valid rule in D.

The proof of ($\Leftarrow$) can be directly obtained from Theorem 1, Theorem 3, and the above definition.

So, $A \rightarrow B$ can be extracted as an approximate rule in the above model if and only if $A \rightarrow B$ is a valid rule in D. □

6 Conclusions

Mining association rules is an expensive process. Mining approximate association rules on a sample of a large database can reduce the computation cost significantly. Srikant and Agrawal [11] suggested a method to select the sample of a given large database for estimating the support of candidates using Chernoff bounds. Also, Toivonen [12] applied the Chernoff bounds to discover association rules in large databases. However, previous approximate models based on Chernoff bounds may require a large sample size compared to the central limit theorem for discovering association rules in large databases. In this paper, we have addressed the issue of mining association rules, and have made the following contributions:

(1) Presented a method of applying the theorems to estimate the size of random database that enables us to mine approximate association rules.

(2) Proposed the algorithm to discover approximate association rules with negative itemsets. In particular, an algorithm of generating all possible useful (positive and negative large) itemsets is also presented.

(3) Demonstrated the effectiveness of our approach experimentally. Our results show that the approximating model is more efficient than models based on Chernoff bounds [11, 12].

(4) Proposed a method to (a) assist knowledge discovery and (b) determine the validation of a rule in the neighbour of min_{supp} or the neighbour of min_{conf} in the given database.

Acknowledgements

The authors would like to thank two anonymous reviewers for their detailed constructive comments on the first version of this paper.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, Mining association rules between sets of items in large databases. In: *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1993:207-216.

[2] R. Agrawal, R. Srikant, Fast algorithms for mining association rules. In: *Proceedings of the 20th VLDB Conference*, 1994:487-499.

[3] S. Brin, R. Motwani and C. Silverstein, Beyond Market Baskets: Generalizing Association Rules to Correlations. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1997: 265-276.

[4] S. Brin, R. Motwani, J. Ullman and S. Tsur, Dynamic Itemset Counting and Implication Rules for Market Basket data. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1997: 255-264.

[5] R. Durrett, Probability: Theory and Examples, *Duxbury Press*, 1996.

[6] E. Omiecinski and A. Savasere, Efficient mining of association rules in large dynamic databases. In: *Proceedings of 16th British National Conference on Databases BNCOD 16*, Cardiff, Wales, UK, 1998.7: 49-63.

[7] J. Park, M. Chen, and P. Yu, Using a Hash-based Method with Transaction Trimming for Mining Association Rules. *IEEE Trans. Knowledge and Data Eng.*, vol. 9, 5(1997): 813-824.

[8] G. Piatetsky-Shapiro, Discovery, analysis, and presentation of strong rules. In: *Knowledge discovery in Databases*, G. Piatetsky-Shapiro and W. Frawley (Eds.), AAAI Press/MIT Press, 1991: 229-248.

[9] A. Savasere, E. Omiecinski, and S. Navathe, An efficient algorithm for mining association rules in large databases. *Proceedings of the 21st International Conference on Very Large Data Bases.* Zurich, Switzerland, 1995.8: 688-692.

[10] T. Shintani and M. Kitsuregawa, Parallel mining algorithms for generalized association rules with classification hierarchy. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1998: 25-36.

[11] R. Srikant and R. Agrawal, Mining generalized association rules. *Future Generation Computer Systems*, Vol. 13, 1997: 161-180.

[12] H. Toivonen, Sampling large databases for association rules. *Proceedings of the 22nd VLDB Conference*, 1996: 134-145.

[13] Shichao Zhang, Aggregation and maintenance for databases mining. *Intelligent Data Analysis: an international journal*, Vol. 3(6) 1999: 475-490.

[14] Shichao Zhang and Xindong Wu, Large Scale Data Mining Based on Data Partitioning. *Applied Artificial Intelligence: an international journal*, Vol. 15 2(2001): 129-139.

[15] Chengqi Zhang and Shichao Zhang, *Association Rules Mining: Models and Algorithms*. Springer, LNAI 2307, p.243, 2002.

[16] Shichao Zhang and Chengqi Zhang, Anytime Mining for Multi-User Applications. *IEEE Transactions on Systems, Man and Cybernetics* (Part B), Vol. 32 No. 4(2002).

Appendix A

The *size* of the given database is 100000 and $min_{supp} = 0.2$. In order to illustrate the efficiency of our approximate rule model, we partly list the experimental results, which are the large itemsets and their supports. The variables a, b, and x_0 are the initialized values used in the random number generator. In order to test the approximation, we list three different supports of each itemset from different samples with the same size as follows.

Some itemsets of PL and NL in original database

```
  /* 1-items */
 Item = China, count = 37000, support = 37%
 Item = Singapore, count = 50000, support = 50%
 Item = B, count = 33200, support = 33.2%
 Item = C, count = 42050, support = 42.05%

/* 2-items */
 Itemset = {China, B}, count = 27750, support = 27.75%
 Itemset = {Singapore, C}, count = 35000, support = 35%
 Itemset = {China, Singapore}, count = 0, support = 0%
 Itemset = {China, C}, count = 3100, support = 3.1%
 Itemset = {B, C}, count = 0, support = 0%
```

Some itemsets of *PL* and *NL* in models based on Chernoff bounds

```
/* parameter value */
  Accuracy of result: 0.01
  Probability of requirements: 0.9
  RandomDBSize: 15000
  a = 53
  b = 113
  x0= 17

  /* 1-item */
 Item = China, count = 5517, support = 36.78%
 Item = Singapore, count = 7565, support = 50.43%
 Item = B, count = 5015, support = 33.43%
 Item = C, count = 6326, support = 42.17%

/* 2-items */
Itemset = {China, B}, count = 4175, support = 27.83%
Itemset = {Singapore, C}, count = 5246, support = 34.97%
Itemset = {China, Singapore}, count = 0, support = 0%
Itemset = {China, C}, count = 431, support = 2.87%
Itemset = {B, C}, count = 0, support = 0%
----------------------------------------------------------
/* parameter value */
  Accuracy of result: 0.01
  Probability of requirements: 0.9
  RandomDBSize: 15000
  a = 53
  b = 113
  x0= 43

  /* 1-item */
 Item = China, count = 5543, support = 36.95%
 Item = Singapore, count = 7449, support = 49.66%
 Item = B, count = 4946, support = 32.97%
 Item = C, count = 6300, support = 42%

/* 2-items */
Itemset = {China, B}, count = 4109, support = 27.39%
Itemset = {Singapore, C}, count = 5249, support = 34.99%
Itemset = {China, Singapore}, count = 0, support = 0%
Itemset = {China, C}, count = 411, support = 2.74%
Itemset = {B, C}, count = 0, support = 0%
```

```
-----------------------------------------------------------
/* parameter value */
  Accuracy of result: 0.01
  Probability of requirements: 0.9
  RandomDBSize: 15000
  a = 53
  b = 113
  x0= 97

  /* 1-item */
 Item = China, count = 5513, support = 36.75%
 Item = Singapore, count = 7568, support = 50.45%
 Item = B, count = 5012, support = 33.41%
 Item = C, count = 6332, support = 42.21%

/* 2-items */
Itemset = {China, B}, count = 4172, support = 27.81%
Itemset = {Singapore, C}, count = 5252, support = 35.01%
Itemset = {China, Singapore}, count = 0, support = 0%
Itemset = {China, C}, count = 473, support = 3.15%
Itemset = {B, C}, count = 0, support = 0%
```

Some itemsets of *PL* and *NL* in central limit theorem

```
/* parameter value */
  Accuracy of result: 0.01
  Probability of requirements: 0.9
  RandomDBSize: 6724
  a = 53
  b = 113
  x0= 17

  /* 1-item */
 Item = China, count = 2453, support = 36.48%
 Item = Singapore, count = 3417, support = 50.82%
 Item = B, count = 2249, support = 33.45%
 Item = C, count = 2844, support = 42.3%

/* 2-items */
Itemset = {China, B}, count = 1863, support = 27.71%
Itemset = {Singapore, C}, count = 2358, support = 35.07%
Itemset = {China, Singapore}, count = 0, support = 0%
Itemset = {China, C}, count = 202, support = 3.01%
Itemset = {B, C}, count = 0, support = 0%
-----------------------------------------------------------
```

```
/* parameter value */ }
  Accuracy of result: 0.01
  Probability of requirements: 0.9
  RandomDBSize: 6724
  a = 53
  b = 113
  x0= 43

  /* 1-item */
 Item = China, count = 2468, support = 36.7%
 Item = Singapore, count = 3350, support = 49.82%
 Item = B, count = 2216, support = 32.96%
 Item = C, count = 2829, support = 42.07%

/* 2-items */
Itemset = {China, B}, count = 1830, support = 27.22%
Itemset = {Singapore, C}, count = 2359, support = 35.08%
Itemset = {China, Singapore}, count = 0, support = 0%
Itemset = {China, C}, count = 196, support = 2.91%
Itemset = {B, C}, count = 0, support = 0%
----------------------------------------------------------
/* parameter value */
  Accuracy of result: 0.01
  Probability of requirements: 0.9
  RandomDBSize: 6724
  a = 53
  b = 113
  x0= 97

  /* 1-item */
 Item = China, count = 2456, support = 36.53%
 Item = Singapore, count = 3412, support = 50.74%
 Item = B, count = 2255, support = 33.54%
 Item = C, count = 2837, support = 42.19%

/* 2-items */
Itemset = {China, B}, count = 1867, support = 27.77%
Itemset = {Singapore, C}, count = 2350, support = 34.95%
Itemset = {China, Singapore}, count = 0, support = 0%
Itemset = {China, C}, count = 204, support = 3.04%
Itemset = {B, C}, count = 0, support = 0%
```

Received October, 2000