Reliable Representations for Association Rules

Yue Xu, Yuefeng Li, Gavin Shaw

Computer Science Discipline
Faculty of Science and Technology
Queensland University of Technology
Brisbane, Australia

Abstract

Association rule mining has contributed to many advances in the area of knowledge discovery. However, the quality of the discovered association rules is a big concern and has drawn more and more attention recently. One problem with the quality of the discovered association rules is the huge size of the extracted rule set. Often for a dataset, a huge number of rules can be extracted, but many of them can be redundant to other rules and thus useless in practice. Mining non-redundant rules is a promising approach to solve this problem. In this paper, we first propose a definition for redundancy, then propose a concise representation, called a Reliable basis, for representing non-redundant association rules. The Reliable basis contains a set of non-redundant rules which are derived using frequent closed itemsets and their generators instead of using frequent itemsets that are usually used by traditional association rule mining approaches. An important contribution of this paper is that we propose to use the certainty factor as the criterion to measure the strength of the discovered association rules. Using this criterion, we can ensure the elimination of as many redundant rules as possible without reducing the inference capacity of the remaining extracted non-redundant rules. We prove that the redundancy elimination, based on the proposed Reliable basis, does not reduce the strength of belief in the extracted rules. We also prove that all association rules, their supports and confidences, can be retrieved from the Reliable basis without accessing the dataset. Therefore the Reliable basis is a lossless representation of association rules. Experimental results show that the proposed Reliable basis can significantly reduce the number of extracted rules. We also conduct experiments on the application of association rules to the area of product recommendation. The experimental results show that the non-redundant association rules extracted using the

proposed method retain the same inference capacity as the entire rule set. This result indicates that using non-redundant rules only is sufficient to solve real problems needless using the entire rule set.

Key words: Knowledge discovery, Association rule mining, Redundant Association rules, Closed itemsets, Data mining agents

---

1. Introduction

For most of the work done in the area of association rule mining, the primary focus has been on developing novel algorithms to aid efficient computation of such rules [AS94, Bay98, HP00], especially regarding improving the efficiency of generating frequent itemsets. However, the quality of the extracted rules has not drawn adequate attention. One big problem in association rule mining is the huge number of extracted rules which results in difficulties in end users' comprehension, and therefore effective usage, of the discovered rules, significantly reducing the effectiveness of association rule mining. If the extracted knowledge can't be effectively used in solving real world problems, the effort used to extract the knowledge is wasted. Moreover, many of the extracted rules produce no value to the user or can be replaced by other rules thus can be considered redundant. The extent of redundancy is much larger than previously suspected [Zak04], especially for dense datasets [PTB+05]. Some efforts have been made to reduce the number of the extracted rules. The approaches can be roughly divided into two categories, subjective approach and objective approach. In the subjective approach category, one technique is to define various interestingness measures and only the rules which are considered of interest based on the interesting measures are generated [BL97, BMUT97]. Another technique in this category is to apply constraints or templates to generate only those rules that satisfy the constraints or templates [MSMG04, DD09, BAG00, HLN99, NLHP98, SVA97].

In the objective approach category, the main technique is to construct concise representations of association rules without applying user-dependent constraints. A concise representation contains a much smaller number of rules and would be considered lossless if all association rules can be derived from the representation. Association rules are derived from frequent itemsets. For a large dataset, especially for a dense dataset where data is heavily correlated, the number of frequent itemsets is often so huge that generating all frequent itemsets requires unrealistic resources (memory and time) [CRB06].

One solution to this problem is to design concise representations of frequent itemsets. A concise representation of frequent itemsets, also called a perfect cover of frequent itemsets in [CCL05], is a proper subset of frequent itemsets, from which all frequent itemsets and their supports can be derived without any further access to the dataset. In the past decade, a number of concise representations of frequent itemsets have been proposed such as closed itemsets [PBTL99a], free itemsets [BBR03], disjunction-free itemsets [BR03], non-derivable itemsets [CG02], and essential itemsets [CCL05]. Initially, the primary purpose of these condensed representations is the efficiency of generating all frequent itemsets [CRB06] rather than for concisely representing association rules. However, among these proposed condensed representations, closed itemsets are of particular interest as they can be applied to generate a condensed set of association rules [KRG04, PBTL99b, Zak00].

The notion of closed frequent itemsets has its origins in the mathematical theory of Formal Concept Analysis introduced in the early 1980s [GW99, Wil82]. An itemset is said to be closed if and only if no proper superset of that itemset has the same support as that itemset. For a given support threshold, knowing all frequent closed itemsets is sufficient to generate all the frequent itemsets and their supports without accessing the dataset. The use of frequent closed itemsets presents a clear promise to reduce the number of extracted rules and also provides a concise representation of association rules [PTB$^+$05, Zak04]. Even though the number of extracted rules can be reduced drastically by only using frequent closed itemsets, a considerable amount of redundancy still remains. Our work will be in this category to construct concise representations of association rules based on closed itemsets for effective redundancy reduction.

We argue that the evaluation of non-redundant rule mining algorithms should take three factors into consideration: removing as much redundancy as possible; ensuring the extracted non-redundant rules retain the same inference capacity; and retaining the ability to retrieve all association rules. In this paper, our goal is to develop techniques that can generate as few rules as possible without reducing the inference capacity of the remaining rules and also without losing any information. To achieve this, we propose a definition of redundant association rules based on which non-redundant association rules can be generated. The non-redundant rules defined in this paper have minimal antecedents and maximal consequents which are similar to the non-redundant rules defined in [PTB$^+$05]. However, our definition relaxes or reduces the requirements to redundancy and thus a much greater

number of redundant rules can be eliminated compared to the approach proposed in [PTB+05]. We propose a concise representation of association rules, called Reliable basis which contains a set of non-redundant rules that meet the proposed definition. Most importantly, in this paper, we propose to use the Certainty Factor (CF) as the criterion to measure the strength of the discovered association rules. The certainty factor is an important and popularly used measure of belief in inference rules [SB75]. With this criterion, we can ensure the removal of the maximal amount of redundancy without reducing the inference capacity of the remaining extracted non-redundant rules. We prove that the redundant rules eliminated by our approach have less or equal CF belief values than that of their corresponding extracted non-redundant rules, and thus that the elimination of such redundant rules will not reduce the belief of the extracted rules. We also show by experiments that the proposed Reliable basis can retain the same or better capacity as the entire rule set to solve problems. Moreover, we prove that the Reliable basis is a lossless representation of association rules since all association rules can be retrieved from the Reliable basis. The contributions of this paper are summarized below:

- We propose a definition of redundant association rules with a relaxing requirement to redundancy so that more redundant rules can be eliminated.
- We propose a concise representation, called Reliable basis, to represent the non-redundant association rules defined in this paper.
- We propose to use the Certainty Factor as the criterion to measure the strength of association rules. We prove that eliminating the redundant rules defined in this paper will not reduce the strength of the extracted non-redundant rules (i.e., Theorem 1).
- We prove that the proposed Reliable basis can be generated from frequent closed itemsets and their generators (i.e., Theorem 2).
- We prove that all association rules can be retrieved from the Reliable basis (i.e., Theorem 3 and Theorem 4). Therefore, the Reliable basis is a lossless representation of association rules.

- We show by experiments that the non-redundant rules contained in the proposed Reliable basis can be used to solve real problems with the same or better capacity as the entire rule set.

The paper is organized as follows. The basic concepts of association rule mining are given in Section 2. In Section 3, we propose a definition of redundancy and then discuss the elimination of the redundancy. Section 4 introduces the proposed Reliable association rule basis for extracting non-redundant rules, then presents a method to retrieve all association rules from the Reliable basis. Experimental results are given in Section 5. Section 6 discusses some related work. Finally, Section 7 concludes the paper.

2. Problem Definition

Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of m distinct items, t be a transaction that contains a set of items such that $t \subseteq I$, T be a dataset containing different identifiable transactions. An association rule is an implication in the form of $X \Rightarrow Y$, where $X, Y \subset I$ are sets of items called itemsets, and $X \bigcap Y = \phi$. The definition of closed itemsets comes from the closure operation of the Galois connection [GW99]. $\forall i \in I$ and $\forall t \in T$, if item i appears in transaction t, then i and t has a binary relation δ denoted as $i\delta t$. The Galois connection of the binary relation is defined by the following mappings where $X \subseteq I$, $Y \subseteq T$:

$$\tau : 2^I \rightarrow 2^T, \tau(X) = \{t \in T | \forall i \in X, i\delta t\} \quad (1)$$

$$\gamma : 2^T \rightarrow 2^I, \gamma(Y) = \{i \in I | \forall t \in Y, i\delta t\} \quad (2)$$

$\tau(X)$ is called the transaction mapping of X. $\gamma(Y)$ is called the item mapping of Y. $\gamma \circ \tau(X)$, called the closure of X, gives the common items among the transactions each of which contains X. With the mappings and the closure defined above, we can formally define the following important concepts in association rule mining.

Definition 1. *(Support) The support of an itemset X, denoted as supp(X), is the percentage of the transactions which contain X, i.e., supp$(X) = |\tau(X)|/|T|$.*

Definition 2. *(Confidence) The confidence of an association rule $X \Rightarrow Y$, denoted as $conf(X \Rightarrow Y)$, is the percentage of the transactions which contain $X \bigcup Y$ out of the transactions which contain X only, i.e., $conf(X \Rightarrow Y) = |\tau(X \bigcup Y)|/|\tau(X)|$.*

Definition 3. *(Closed Itemset) Let X be a subset of I. X is a frequent closed itemset iff $\gamma \circ \tau(X) = X$.*

Definition 4. *(Generator) An itemset $g \in 2^I$ is a generator of a closed itemset $c \in 2^I$ iff $c = \gamma \circ \tau(g)$ and $g \subset \gamma \circ \tau(g)$. g is said to be a minimal generator of the closed itemset set c if $\nexists g' \subset g$ such that $\gamma \circ \tau(g') = c$.*

From Definition 4, we can get that $g \subset c$ is true for any generator g and its closed itemset c. The Galois connection satisfies the following properties[GW99].

Property 1. *Let X, $c \in 2^I$. If c is the closed itemset of X, then $supp(X) = supp(c)$.*

Property 2. *Let X, X_1, $X_2 \in 2^I$ and Y, Y_1, $Y_2 \in 2^T$.*

1. $X_1 \subseteq X_2 \Longrightarrow \tau(X_1) \supseteq \tau(X_2)$
2. $Y_1 \subseteq Y_2 \Longrightarrow \gamma(Y_1) \supseteq \gamma(Y_2)$

Property 2 indicates that the transaction mapping of an itemset is larger than the transaction mapping of its super itemset, and that the common itemset of a transaction set (i.e., the item mapping of the transaction set) is larger than the common itemset of a super set of the transaction set. These properties reflect the nature of itemsets and will be used in theorem proofs in the following sections.

A few examples are given below to illustrate the concepts defined above. The following simple dataset involves 5 items and consists of 6 transactions, i.e., I=$\{A, B, C, D, E\}$ and T=$\{1, 2, 3, 4, 5, 6\}$ where the transactions are identified using their ID numbers.

For itemsets AC and $ABCE$, their transaction mappings are $\tau(AC) = 135$ and $\tau(ABCE) = 35$, respectively. The item mapping of transaction set $\{3, 5\}$ is $\gamma(35) = ABCE$. The support of AC and $ABCE$ is 1/2 and 1/3, respectively and the confidence of rule $AC \Rightarrow BE$ is 2/3.

Transaction ID	Items
1	A C D
2	B C E
3	A B C E
4	B E
5	A B C E
6	B C E

Closed itemsets	Minimal Generators
AC	A
BE	B, E
BCE	BC, CE
ABCE	AB, AE

From the simple dataset, we can generate the following closed itemsets and corresponding minimal generators:

Association rule mining is usually decomposed into two sub problems: to find frequent itemsets whose support is larger than or equal to the predefined minimum support and then from those frequent itemsets to generate association rules that satisfy the minimum support and minimum confidence. For the popular Mushroom dataset (http://kdd.ics.uci.edu/), with minimum support 0.8 and minimum confidence 0.8, we can generate 88 association rules. Table 1 displays 20 of the 88 association rules. The closed itemsets and their minimal generators are given in Table 2.

3. Redundancy in Association Rules

A challenge to association mining is the huge number of extracted rules. Recent studies have shown that using closed itemsets and generators to extract association rules can greatly reduce the number of extracted rules [PTB$^+$05, Zak00]. However, a considerable amount of redundancy still exists in the rules extracted based on closed itemsets. Therefore, techniques are needed to remove the redundancy in order to generate high quality association rules. The scope of the redundancy must be carefully and fairly defined so that the reduction won't cause information loss or reduce the belief in the

Table 1: Association rules (Mushroom Dataset, minsupp=0.8, minconf=0.8)

	Rules (supp, conf)
1	gill-attachment-f ⇒ veil-type-p (0.97415,1.0)
2	veil-color-w ⇒ veil-type-p (0.97538,1.0)
3	gill-attachment-f,veil-color-w ⇒ veil-type-p (0.97317,1.0)
4	gill-attachment-f,ring-number-o ⇒ veil-type-p (0.89808,1.0)
5	gill-spacing-c,veil-color-w ⇒ veil-type-p (0.81487,1.0)
6	gill-attachment-f,gill-spacing-c ⇒ veil-type-p,veil-color-w (0.81265,1.0)
7	gill-attachment-f,gill-spacing-c⇒ veil-type-p (0.81265,1.0)
8	gill-attachment-f,gill-spacing-c,veil-type-p ⇒ veil-color-w (0.81265,1.0)
9	gill-attachment-f ⇒ veil-type-p,veil-color-w (0.97317,0.99899)
10	gill-attachment-f ⇒ veil-type-p,ring-number-o (0.89808,0.92191)
11	veil-color-w ⇒ gill-spacing-c,veil-type-p (0.81487,0.83544)
12	veil-color-w ⇒ gill-attachment-f,gill-spacing-c,veil-type-p (0.81265,0.83317)
13	gill-attachment-f,veil-color-w ⇒ gill-spacing-c,veil-type-p (0.81265,0.83506)
14	gill-attachment-f,veil-color-w ⇒ veil-type-p,ring-number-o (0.8971,0.92183)
15	gill-attachment-f,ring-number-o ⇒ veil-type-p,veil-color-w (0.8971,0.9989)
16	gill-spacing-c,veil-color-w ⇒ gill-attachment-f,veil-type-p (0.81265,0.99728)
17	gill-attachment-f ⇒ veil-color-w (0.97317,0.99899)
18	gill-attachment-f ⇒ ring-number-o (0.89808,0.92191)
19	gill-attachment-f,veil-color-w ⇒ gill-spacing-c (0.81265,0.83506)
20	gill-attachment-f,ring-number-o ⇒ veil-color-w (0.8971,0.9989)

Table 2: Closed Itemsets and Minimal Generators (Mushroom Dataset, minsupp=0.8)

Closed itemsets	**Minimal Generators**	**Support**
{veil-type-p}		1.0
{gill-attachment-f,veil-type-p}	{gill-attachment-f}	0.97415
{gill-spacing-c,veil-type-p}	{gill-spacing-c}	0.8385
{veil-type-p,veil-color-w}	{veil-color-w}	0.97538
{veil-type-p,ring-number-o}	{ring-number-o}	0.9217
{gill-attachment-f, veil-type-p,veil-color-w}	{gill-attachment-f, veil-color-w}	0.97317
{gill-attachment-f,veil-type-p, ring-number-o}	{gill-attachment-f, ring-number-o}	0.8981
{gill-spacing-c,veil-type-p, veil-color-w}	{gill-spacing-c, veil-color-w}	0.81487
{gill-attachment-f,gill-spacing-c, veil-type-p,veil-color-w}	{gill-attachment-f, gill-spacing-c}	0.81265
{gill-attachment-f,veil-type-p, veil-color-w,ring-number-o}	{veil-color-w, ring-number-o}	0.8971

resulting rules. Any information loss or belief degradation will cause quality deterioration of the extracted rules, which makes the redundancy reduction not worthwhile. In this section, we start with some examples to show the existence of redundancy in association rules, following that we give a definition of redundant rules, and then we prove that the elimination of the defined redundancy won't reduce the belief in the extracted non-redundant rules. In Section 4, we describe a concise representation of the defined non-redundant association rules, from which all association rules can be derived.

3.1. Redundancy Definition

The rules in Table 1 are considered useful based on the predefined minimum support and confidence. However, some of the rules actually do not contribute new information. The consequent concluded by some rules can be obtained via other rules with higher or the same confidence but without requiring more conditions to be satisfied. For example, in order to be selected the rules 5, 8, 13, and 20 in Table 1 require more conditions to be satisfied than that of rules 2, 6, 11, and 9, respectively, but conclude the same or less results which can be produced by rules 2, 6, 11, and 9. That means, without

rules 5, 8, 13, and 20, we still can achieve the same result using other rules. Therefore, rules 5, 8, 13, and 20 are considered redundant to rules 2, 6, 11, and 9, respectively. Compared to rules 2, 6, 11, and 9, the redundant rules 5, 8, 13, and 20 have a longer or the same antecedent and a shorter or the same consequent, respectively, and the confidence of the redundant rules is not larger than that of their corresponding non-redundant rules. The following definition defines this type of redundant rules.

Definition 5. *(Redundant rules) Let $X \Rightarrow Y$ and $X' \Rightarrow Y'$ be two association rules with confidence cf and cf', respectively. $X \Rightarrow Y$ is said to be a redundant rule to $X' \Rightarrow Y'$ if $X' \subseteq X$, $Y \subseteq Y'$, and $cf \leq cf'$.*

Based on Definition 5, for an association rule $X \Rightarrow Y$, if there does not exist any other rule $X' \Rightarrow Y'$ such that the confidence of $X' \Rightarrow Y'$ is the same as or larger than the confidence of $X \Rightarrow Y$, $X' \subseteq X$ or $Y \subseteq Y'$, then $X \Rightarrow Y$ is non-redundant. In terms of the requirement of shorter antecedent and longer consequent, Definition 5 is similar to the definition of min-max association rules defined in [PTB+05]. However, the definition of min-max association rules requires that a redundant rule and its corresponding non-redundant rule must have identical confidence and identical support, while Definition 5 here only requires that the confidence of the redundant rule is not larger than that of its corresponding non-redundant rule.

Safely eliminating redundancy without damaging the capacity of the remaining rules is an essential issue and defining a boundary between redundancy and non-redundancy is crucial to ensure safe redundancy elimination. Several approaches have been proposed for redundancy elimination [PBTL99b, Zak00, CG02]. However, none of them have specifically discussed the boundary. As mentioned above, in this paper, we propose to use Certainty Factor (CF) defined in Section 3.2 as a criterion to determine the boundary. If the deletion of a rule does not reduce the CF value of the remaining rules, the deletion is considered safe. In the following subsection, we prove that the elimination of the redundancy defined by Definition 5 will not reduce the belief in the remaining non-redundant rules.

3.2. Reliable Redundancy Elimination

The certainty factor theory was first introduced in MYCIN [SB75] to express how accurate and truthful a rule is and how reliable the antecedent of the rule is. MYCIN was an early expert system developed at Stanford

University in the early 1970s for the diagnosis of blood clotting diseases. The certainty factor theory is based on two functions: measure of belief $MB(X,Y)$ and measure of disbelief $MD(X,Y)$ for a rule $X \Rightarrow Y$, as given below.

$$MB(X,Y) = \begin{cases} 1 & P(Y) = 1 \\ 0 & P(Y/X) \leq P(Y) \\ \frac{P(Y/X)-P(Y)}{1-P(Y)} & otherwise \end{cases} \quad (3)$$

$$MD(X,Y) = \begin{cases} 1 & P(Y) = 0 \\ 0 & P(Y/X) \geq P(Y) \\ \frac{P(Y)-P(Y/X)}{P(Y)} & otherwise \end{cases} \quad (4)$$

where, in the context of association rules, $P(Y/X)$ and $P(Y)$ are the confidence of the rule and the support of the consequent, respectively. The values of $MB(X,Y)$ and $MD(X,Y)$ range between 0 and 1 measuring the strength of belief or disbelief in consequent Y given antecedent X. $MB(X,Y)$ weighs how much the antecedent X increases the possibility of Y occurring. Similarly, $MD(X,Y)$ weighs how much the antecedent X decreases the possibility of Y occurring. If the antecedent completely supports the consequent, then $P(Y/X)$ will be equal to 1 thus $MB(X,Y)$ will be 1. If $P(Y/X)$=0 which indicates that the antecedent completely denies the consequent, then $MD(X,Y)$=1 thus the disbelief in the rule reaches its highest value. The total strength of belief or disbelief in the association captured by the rule is measured by the certainty factor which is defined as follows:

$$CF(X,Y) = MB(X,Y) - MD(X,Y) \quad (5)$$

The value of a certainty factor is between 1 and -1. Negative values represent cases where the antecedent is against the consequent; positive values represent cases where the antecedent supports the consequent; while CF=0 means that the antecedent does not influence the belief in Y. Obviously, association rules with high CF values are more useful since they represent strong positive associations between antecedents and consequents. Indeed, the aim of association rule mining is to discover strong positive associations from large amounts of data. Therefore, we propose that the certainty factor can be used to measure the strength of discovered association rules.

Theorem 1 below states that the CF value of a redundant rule defined by Definition 5 will never be larger than the CF value of its corresponding non-redundant rules. It means that, the association between the antecedent and consequent of the non-redundant rule is stronger than that of the corresponding redundant rule.

Theorem 1. *Let $X \Rightarrow Y$ and $X' \Rightarrow Y'$ be two association rules. If $Y' \subseteq Y$, and $P(Y/X) \geq P(Y'/X')$, then $CF(X,Y) \geq CF(X',Y')$.*

PROOF. We can prove the theorem, i.e., $CF(X,Y) \geq CF(X',Y')$, by proving that $CF(X,Y)$ - $CF(X',Y') \geq 0$.
From Equation (5) we have $CF(X,Y)$ - $CF(X',Y') = MB(X,Y)$ - $MB(X',Y')$ + $MD(X',Y')$-$MD(X,Y)$. Hence, we need to prove that $MB(X,Y)$ - $MB(X',Y')$ + $MD(X',Y')$-$MD(X,Y) \geq 0$.

1. Assuming that $P(Y'/X') \geq P(Y')$. From condition $Y' \subseteq Y$, we have $P(Y) \leq P(Y')$. Because $P(Y/X) \geq P(Y'/X')$, we have $P(Y/X) \geq P(Y)$. Therefore in this case, $MD(X',Y')$ - $MD(X,Y) = 0$. To prove the theorem, we need to prove that $MB(X,Y)$ - $MB(X',Y') \geq 0$. From Equation (3), we have:
   $MB(X,Y)$ - $MB(X',Y') = \frac{P(Y/X)-P(Y)}{1-P(Y)} - \frac{P(Y'/X')-P(Y')}{1-P(Y')}$
   $= \frac{P(Y/X)-P(Y'/X')+P(Y'/X')P(Y)-P(Y/X)P(Y')-P(Y)+P(Y')}{(1-P(Y))(1-P(Y'))}$
   $= \frac{(P(Y/X)-P(Y'/X'))(1-P(Y'))+(P(Y')-P(Y))(1-P(Y'/X'))}{(1-P(Y))(1-P(Y'))}$

   Because $P(Y) \leq P(Y')$ and $P(Y/X) \geq P(Y'/X')$, we prove that the above expression ≥ 0. Hence, $MB(X,Y)$ - $MB(X',Y') \geq 0$
2. Assuming that $P(Y'/X') \leq P(Y')$. In this situation, we have two cases.

   (a) $P(Y/X) \leq P(Y)$
   In this case, $MB(X,Y)$ - $MB(X',Y') = 0$. To prove the theorem, we need to prove that $MD(X',Y')$ - $MD(X,Y) \geq 0$. From Equation (4), we have
   $MD(X',Y')$ - $MD(X,Y) = \frac{P(Y')-P(Y'/X')}{P(Y')} - \frac{P(Y)-P(Y/X)}{P(Y)}$
   $= \frac{P(Y/X)P(Y')-P(Y'/X')P(Y)}{P(Y)P(Y')} \geq \frac{P(Y/X)P(Y')-P(Y/X)P(Y)}{P(Y)P(Y')}$.
   Again, since $P(Y) \leq P(Y')$, we get $MD(X',Y')$ - $MD(X,Y) \geq 0$.

   (b) $P(Y/X) \geq P(Y)$
   In this case, $MD(X,Y)=0$ and $MB(X',Y') = 0$. To prove the theorem, we need to prove $MD(X',Y') + MB(X,Y) \geq 0$. Because $P(Y'/X') \leq P(Y')$ and $P(Y/X) \geq P(Y)$, from the equations (3) and (4), it is true that $MD(X',Y') + MB(X,Y) \geq 0$
   Combining the results of the above cases, we have $CF(X,Y)$ - $CF(X',Y') \geq 0$, hence $CF(X,Y) \geq CF(X',Y')$.

□

Theorem 1 states that, as long as $Y' \subseteq Y$ (no matter whether $X \subseteq X'$ or not) and $P(Y/X) \geq P(Y'/X')$, i.e., the confidence of $X \Rightarrow Y$ is not less than the confidence of $X' \Rightarrow Y'$, the CF value of $X \Rightarrow Y$ will not be less than that of $X' \Rightarrow Y'$. According to Definition 5, if $Y' \subseteq Y$, $X \subseteq X'$, and

the confidence of $X \Rightarrow Y$ is not less than the confidence of $X' \Rightarrow Y'$, then $X' \Rightarrow Y'$ is considered redundant to $X \Rightarrow Y$. This means that, based on Theorem 1, the CF value of the redundant rule $X' \Rightarrow Y'$ is never higher than that of its corresponding non-redundant rule $X \Rightarrow Y$ and thus the elimination of $X' \Rightarrow Y'$ is reliable since it won't reduce the belief in the extracted non-redundant rule $X \Rightarrow Y$.

4. Concise Bases Representing Non-redundant Association Rules

Developing concise and lossless representations is a promising way to improve the quality of the discovered associations. Some work has been done in this area [GMT05, KRG04, PTB$^+$05, Zak04]. Pasquier et al. [PTB$^+$05] proposed two condensed bases to represent non-redundant association rules, which are defined as follows:

Definition 6. *(Min-max Approximate Basis) Let C be the set of frequent closed itemsets and G be the set of minimal generators of the frequent closed itemsets in C. The min-max approximate basis is:*

$$MinMaxApprox = \{g \Rightarrow (c \backslash g) | c \in C, g \in G, \gamma \circ \tau(g) \subset c\}$$

Definition 7. *(Min-max Exact Basis) Let C be the set of frequent closed itemsets. For each frequent closed itemset c, let G_c be the set of minimal generators of c. The min-max exact basis is:*

$$MinMaxExact = \{g \Rightarrow (c \backslash g) | c \in C, g \in G_c, g \neq c\}$$

Rules with confidence less than 1 are called Approximate rules and rules with confidence equal to 1 are called Exact rules. For the 88 rules extracted from the Mushroom dataset mentioned above, there are 17 exact rules and 71 approximate rules. Based on the Min-max approximate basis and the Min-max exact basis, only 9 exact rules and 25 approximate rules, as displayed in Table 3 and Table 4, are extracted and considered non-redundant in terms of the redundancy definition given in [PTB$^+$05]. However, under Definition 5, some of the rules extracted from the min-max bases are redundant such as rules 5, 6 and 7 in Table 3 which are redundant to rules 1 and 2 in the same table, and rules 22 to 25 in Table 4 which are redundant to rules 17, 11, 10, and 16, respectively.

Table 3: Non-redundant Exact Rules Extracted From Min-max Exact Basis (Mushroom Dataset, minsupp=0.8, minconf=0.8)

	Rules (supp, conf)
1	gill-attachment-f ⇒ veil-type-p (0.97415,1.0)
2	gill-spacing-c ⇒ veil-type-p (0.8385,1.0)
3	veil-color-w ⇒ veil-type-p (0.97538,1.0)
4	ring-number-o ⇒ veil-type-p (0.92171,1.0)
5	gill-attachment-f,veil-color-w ⇒ veil-type-p (0.97317,1.0)
6	gill-attachment-f,ring-number-o ⇒ veil-type-p (0.89808,1.0)
7	gill-spacing-c,veil-colo-w ⇒ veil-type-p (0.81487,1.0)
8	gill-attachment-f,gill-spacing-c ⇒ veil-type-p,veil-color-w (0.81265,1.0)
9	veil-color-w,ring-number-o ⇒ gill-attachment-f,veil-type-p (0.8971,1.0)

4.1. Reliable Bases

Corresponding to the two Min-Max bases, we propose two more concise bases called Reliable bases which are defined in Definition 8 and Definition 9. Using the Reliable bases, more redundant rules can be eliminated.

Definition 8. *(Reliable Approximate Basis) Let C be the set of frequent closed itemsets and G be the set of minimal generators of the frequent closed itemsets in C. The Reliable approximate basis is:*

$$ReliableApprox = \{g \Rightarrow (c \backslash g) | c \in C, g \in G, \gamma \circ \tau(g) \subset c, \neg(g \supseteq ((c \backslash c') \cup g')) \text{ or } conf(g \Rightarrow (c \backslash g)) > conf(g' \Rightarrow (c' \backslash g')) \text{ where } \forall c' \in C, \forall g' \in G, g' \subset g, \gamma \circ \tau(g') \subset c'\}$$

Definition 9. *(Reliable Exact Basis) Let C be the set of frequent closed itemsets. For each frequent closed itemset c, let G_c be the set of minimal generators of c. The Reliable exact basis is:*

$$ReliableExact = \{g \Rightarrow (c \backslash g) | c \in C, g \in G_c, \neg(g \supseteq ((c \backslash c') \cup g')), \text{ where } \forall c' \in C, c' \subset c, \forall g' \in G_{c'}\}$$

It can be proved by the following lemma and theorem that the rules defined by the Reliable bases are non-redundant.

Lemma 1. *Let $c \in C$ and C be the set of frequent closed itemsets, let $g \in G$ and G be the set of minimal generators of the closed itemsets in C. If $\exists c' \in C$, $\exists g' \in G$, $\gamma \circ \tau(g') \subseteq c'$, $g' \subset g$, $g \supseteq ((c \backslash c') \cup g')$, and $conf(g \Rightarrow c \backslash g) \leq conf(g' \Rightarrow c' \backslash g')$, then $g \Rightarrow c \backslash g$ is redundant to $g' \Rightarrow c' \backslash g'$.*

Table 4: Non-redundant Approximate Rules Extracted From Min-max Approximate Basis (Mushroom Dataset, minsupp=0.8, minconf=0.8)

	Rules (supp, conf)
1	veil-type-p ⇒ gill-attachment-f (0.97415,0.97415)
2	veil-type-p ⇒ gill-spacing-c (0.8385,0.8385)
3	veil-type-p ⇒ veil-color-w (0.97538,0.97538)
4	veil-type-p ⇒ ring-number-o (0.92171,0.92171)
5	veil-type-p ⇒ gill-attachment-f,veil-color-w (0.97317,0.97317)
6	veil-type-p ⇒ gill-attachment-f,ring-number-o (0.89808,0.89808)
7	veil-type-p ⇒ gill-spacing-c,veil-color-w (0.81487,0.81487)
8	veil-type-p ⇒ gill-attachment-f,gill-spacing-c, veil-color-w (0.81265,0.81265)
9	veil-type-p ⇒ gill-attachment-f,veil-color-w, ring-number-o (0.8971,0.8971)
10	gill-attachment-f ⇒ veil-type-p,veil-color-w (0.97317,0.99899)
11	gill-attachment-f ⇒ veil-type-p,ring-number-o (0.89808,0.92191)
12	gill-attachment-f ⇒ gill-spacing-c,veil-type-p, veil-color-w (0.81265,0.83422)
13	gill-attachment-f ⇒ veil-type-p,veil-color-w, ring-number-o (0.8971,0.9209)
14	gill-spacing-c ⇒ veil-type-p,veil-color-w (0.81487,0.97181)
15	gill-spacing-c ⇒ gill-attachment-f,veil-type-p, veil-color-w (0.81265,0.96917)
16	veil-color-w ⇒ gill-attachment-f,veil-type-p (0.97317,0.99773)
17	veil-color-w ⇒ gill-spacing-c,veil-type-p (0.81487,0.83544)
18	veil-color-w ⇒ gill-attachment-f,gill-spacing-c,veil-type-p (0.81265,0.83317)
19	veil-color-w ⇒ gill-attachment-f,veil-type-p,ring-number-o (0.8971,0.91974)
20	ring-number-o ⇒ gill-attachment-f,veil-type-p (0.89808,0.97436)
21	ring-number-o ⇒ gill-attachment-f,veil-type-p,veil-color-w (0.8971,0.97329)
22	gill-attachment-f,veil-color-w ⇒ gill-spacing-c,veil-type-p (0.81265,0.83506)
23	gill-attachment-f,veil-color-w ⇒ veil-type-p,ring-number-o (0.8971,0.92183)
24	gill-attachment-f,ring-number-o ⇒ veil-type-p, veil-color-w (0.8971,0.9989)
25	gill-spacing-c,veil-color-w ⇒ gill-attachment-f,veil-type-p (0.81265,0.99728)

PROOF. Let $A = c\backslash c'$ so that $c \subseteq A \cup c'$ and $A \cap c' = \phi$. Therefore, we have $c\backslash((c\backslash c') \cup g') \subseteq (A \cup c')\backslash(A \cup g')$. Since $A \cap c' = \phi$ and $g' \subseteq c'$, then $A \cap g' = \phi$. So, we have
$c\backslash((c\backslash c') \cup g') \subseteq (A \cup c')\backslash(A \cup g') = ((A \cup c')\backslash A)\backslash g') = c'\backslash g'$. That is, $c\backslash((c\backslash c') \cup g') \subseteq c'\backslash g'$. Because $g \supseteq ((c\backslash c') \cup g')$, we have $c\backslash g \subseteq c\backslash((c\backslash c') \cup g') \subseteq c'\backslash g'$, hence, $c\backslash g \subseteq c'\backslash g'$. Since $c\backslash g \subseteq c'\backslash g'$, $g \supset g'$, and $conf(g \Rightarrow c\backslash g) \leq conf(g' \Rightarrow c'\backslash g')$, according to Definition 5, we can conclude that $g \Rightarrow c\backslash g$ is redundant to $g' \Rightarrow c'\backslash g'$.

□

According to Modus tollens inference rule, from Lemma 1, we get the following corollary:

Corollary 1. *Let $c \in C$ and C be the set of frequent closed itemsets, let $g \in G$ and G be the set of minimal generators of the closed itemsets in C, and $\gamma \circ \tau(g) \subseteq c$. If $g \Rightarrow c\backslash g$ is a non-redundant rule, then $\forall c' \in C$, $\forall g' \in G$, $\gamma \circ \tau(g') \subseteq c'$ and $g' \subset g$, we have $\neg(g \supseteq ((c\backslash c') \cup g'))$ or $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$.*

Theorem 2. *Let $c \in C$ and C be the set of frequent closed itemsets, let $g \in G$ and G be the set of minimal generators of the closed itemsets in C, and $\gamma \circ \tau(g) \subseteq c$. $g \Rightarrow c\backslash g$ is a non-redundant rule iff $\forall c' \in C$, $\forall g' \in G$, $\gamma \circ \tau(g') \subseteq c'$, and $\neg(g \supseteq ((c\backslash c') \cup g'))$ or $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$.*

In Theorem 2, if $\gamma \circ \tau(g) = c$ is true, $g \Rightarrow c\backslash g$ is an exact rule, otherwise an approximate rule. Therefore, Theorem 2 covers both Exact basis and Approximate basis.

PROOF.

1. Completeness: if $g \Rightarrow c\backslash g$ is a non-redundant rule, then $\forall c' \in C$, $\forall g' \in G$, $\gamma \circ \tau(g') \subseteq c'$, and $\neg(g \supseteq ((c\backslash c') \cup g'))$ or $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$. Two cases:
   (a) If $g' \subset g$, the proof follows the conclusion of Corollary 1.
   (b) If $g' \supseteq g$, then $g \supseteq ((c\backslash c') \cup g')$ won't be true, i.e., $\neg(g \supseteq ((c\backslash c') \cup g'))$.
2. Soundness: if $\forall c' \in C$, $\forall g' \in G$, $\gamma \circ \tau(g') \subseteq c'$, and $\neg(g \supseteq ((c\backslash c') \cup g'))$ or $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$, then $g \Rightarrow c\backslash g$ is a non-redundant rule.

(a) Assuming that $\neg(g \supseteq ((c\backslash c') \cup g'))$, we get $g \subset (c\backslash c') \cup g'$, or $g \cap ((c\backslash c') \cup g') = \emptyset$, or $(g \cap ((c\backslash c') \cup g') \subset ((c\backslash c') \cup g')) \wedge (g \cap ((c\backslash c') \cup g') \subset g)$.
(1). In the case that $g \subset (c\backslash c') \cup g'$ is true, assuming that $g \Rightarrow c\backslash g$ is redundant, then we get, $\exists c' \in C$, $\exists g' \in G$, and $\gamma \circ \tau(g') \subseteq c'$ (hence $g' \subset c'$) such that $g' \subseteq g$ and $c'\backslash g' \supseteq c\backslash g$.
From $c'\backslash g' \supseteq c\backslash g$ and $g' \subset c'$, we have $c' \supseteq c'\backslash g' \supseteq c\backslash g$. Since $\gamma \circ \tau(g) \subseteq c$ thus $g \subset c$, obviously we have $c = (c\backslash g) \cup g$ and $(c\backslash g) \cap g = \phi$, therefore, we have $c\backslash(c\backslash g) = g$.
Because $c' \supseteq c\backslash g$, hence $c\backslash c' \subseteq c\backslash(c\backslash g) = g$. Therefore, we have $c\backslash c' \subseteq g$. From $g' \subseteq g$, we get $(c\backslash c') \cup g' \subseteq g \cup g' = g$, i.e., $(c\backslash c') \cup g' \subseteq g$ which contradicts to $(c\backslash c') \cup g' \supset g$.
Therefore, the assumption is false, i.e., $g \Rightarrow c\backslash g$ is non-redundant.
(2). In the case that $g \cap ((c\backslash c') \cup g') = \emptyset$ is true, then $g \cap g' = \emptyset$, thus $g \supseteq g'$ is always false. Therefore, $g \Rightarrow c\backslash g$ can't be redundant to $g' \Rightarrow c'\backslash g'$.
(3). In the case that $(g \cap ((c\backslash c') \cup g') \subset ((c\backslash c') \cup g')) \wedge (g \cap ((c\backslash c') \cup g') \subset g)$ is true, there must exist some x such that $x \in c\backslash c'$ and $x \notin g$ or $x \in g'$ and $x \notin g$. The former will make $(c\backslash g) \subset (c'\backslash g')$ false and the latter will make $g \supset g'$ false. Therefore, $g \Rightarrow c\backslash g$ will never be redundant to $g' \Rightarrow c'\backslash g'$.

(b) Assuming that $conf(g \Rightarrow c\backslash g) > conf(g' \Rightarrow c'\backslash g')$. From Definition 5, we can directly conclude that $g \Rightarrow c\backslash g$ is not redundant.

□

Thus we proved that, the rules defined by the Reliable bases are non-redundant. According to our definition of redundant rules given in Definition 5, for two rules $X \Rightarrow Y$ and $X' \Rightarrow Y'$, as long as $X \subseteq X'$, $Y' \subseteq Y$, and the confidence of $X \Rightarrow Y$ is not less than that of $X' \Rightarrow Y'$, $X' \Rightarrow Y'$ is considered redundant to $X \Rightarrow Y$ no matter what the supports of the two rules are. However, for Min-Max basis rules, as proved by Proposition 3 and Proposition 4 in [PTB+05], the support and confidence of a non-redundant rule in the Min-Max basis must be identical to the support and confidence of its corresponding redundant rule, respectively. Our relaxed requirement to redundancy allows more rules to be considered redundant and therefore eliminated. Even though more rules are eliminated based on the Reliable basis, the elimination is reliable because, as explained in Section 3.2, the CF value of a redundant rule is never higher than that of its corresponding

non-redundant rule.

The following property states that the generator of a closed itemset won't be larger than or equal to the generator of its super closed itemset. Therefore, the rules generated from a closed itemset won't be redundant to the rules generated from its super closed itemset. Thus when calculating non-redundant exact rules from a closed itemset c using the Reliable Exact Basis, only sub closed itemsets of c need to be checked. This property is reflected in the definition of Reliable Exact Basis where only subsets $c' \subset c$ are checked.

Property 3. *Let g and g' be minimal generators of c and c', respectively, c and c' be closed itemsets, then $c \subset c' \Rightarrow \neg(g \supseteq g')$.*

PROOF. Assume that $g \supseteq g'$. From Property 2-(1) and Property 2-(2), we get $g \supseteq g' \Rightarrow c \supseteq c'$.
Negating both sides of the above implication by using Modus tolen inference rule, we have $\neg(c \supseteq c') \Rightarrow \neg(g \supseteq g')$. That is, $(c \subset c') \vee (c \cap c' = \emptyset) \vee ((c \cap c' \neq \emptyset) \wedge (c \not\subset c')) \Rightarrow \neg(g \supseteq g')$. Because $(c \subset c')$, $c \cap c' = \emptyset$, and $(c \cap c' \neq \emptyset) \wedge (c \not\subset c')$ are exclusive events, they can't be true simultaneously. Therefore we have $(c \subset c') \Rightarrow \neg(g \supseteq g')$.

□

The generic representation resulting from coupling the Reliable Exact Basis with the Reliable Approximate Basis defines a more concise set of association rules which are non-redundant, sound and lossless. The algorithms to extract non-redundant exact rules and non-redundant approximate rules based on the Reliable bases are given below:

Algorithm 1. ReliableExactRule(*Closure*)
Input: *Closure: a set of frequent closed itemsets*
Output: *A set of non-redundant exact rules.*
1. exactRules $:= \emptyset$
2. for each $c \in Closure$
3. for each $g \in G.c$ *(G.c is the set of minimal generators of c)*
4. if $\forall c' \in Closure$ *such that* $c' \subset c$ *and* $\forall g' \in G.c'$
5. we have $\neg(g \supseteq ((c \backslash c') \cup g'))$
6. then $exactRules := exactRules \cup \{g \Rightarrow (c \backslash g)\}$
7. end
8. end
9. Return exactRules

Algorithm 2. ReliableApproxBasis(*Closure*)

Input: *Closure: a set of frequent closed itemsets*
Generator: a set of minimal generators

Output: *A set of non-redundant approximate rules.*

1. $approxRules := \emptyset$
2. *for each* $c \in Closure$
3. *for each* $g \in Generator$ *such that* $\gamma \circ \tau(g) \subset c$
4. *if* $\forall c' \in Closure$, $\forall g' \in G$ *such that* $\gamma \circ \tau(g') \subset c'$ *and* $g' \subseteq g$
5. *we have* $\neg(g \supseteq ((c \backslash c') \cup g'))$ *or* $conf(g \Rightarrow c \backslash g) > conf(g' \Rightarrow c' \backslash g')$
6. *then* $approxRules := approxRules \cup \{g \Rightarrow (c \backslash g)\}$
7. *end-for*
8. *end-for*
9. *Return appproxRules*

For the Mushroom example dataset, 6 non-redundant exact rules are extracted based on the Reliable Exact basis and 21 non-redundant approximate rules are extracted based on the Reliable Approximate basis. Rules 5, 6, and 7 in Table 3 extracted based on the Min-max Exact basis and rules 22, 23, 24 and 25 in Table 4 extracted based on the Min-max Approximate basis are considered redundant under the Reliable Exact basis and the Reliable Approximate basis, respectively, and thus eliminated.

4.2. Deriving Association Rules from the Reliable Bases

The proposed reliable bases are lossless, which means that from the bases we can construct all association rules without scanning the dataset. Algorithms have been proposed to derive all association rules from the Min-max bases [PTB+05]. In this section, we provide algorithms that can derive all exact and approximate rules from the Reliable bases.

4.2.1. Deriving Exact Association Rules

Based on the definitions 7 and 9, the Min-max exact basis can be described as:

$MinMaxExact = \{g \Rightarrow (c \backslash g) | c \in C, g \in G_c, g \neq c\}$
$= \{g \Rightarrow (c \backslash g) | c \in C, g \in G_c,$

$$\neg(g \supseteq ((c\backslash c') \cup g'))\ for\ all\ c' \in C, c' \subset c, g' \in G_{c'}\} \cup$$
$$\{g \Rightarrow (c\backslash g) | c \in C, g \in G_c,$$
$$g \supseteq ((c\backslash c') \cup g')\ for\ some\ c' \in C, c' \subset c, g' \in G_{c'}\}$$
$$= ReliableExact \cup NonReliableExact$$

Where $NonReliableExact = \{g \Rightarrow (c\backslash g) | c \in C, g \in G_c, g \supseteq ((c\backslash c') \cup g')\ for\ some\ c' \in C, c' \subset c, g' \in G_{c'}\}$

The equation above shows that the Reliable exact basis is a subset of the Min-max exact basis. Pasquier et al. have proved that all exact association rules can be deduced from the Min-max exact basis [PTB+05]. Given $ReliableExact$, if we can deduce $NonReliableExact$ from $ReliableExact$, then we will be able to deduce all exact association rules. The following theorem allows us to generate all NonReliableExact rules from $ReliableExact$.

Theorem 3. *Let C be the set of frequent closed itemsets. For rules $r_1 : g_1 \Rightarrow c_1\backslash g_1$ and $r_2 : g_2 \Rightarrow c_2\backslash g_2$ where c_1, $c_2 \in C$, $g_1 \in G_{c_1}$ and $g_2 \in Gc_2$, r_1 is a NonReliable exact rule iff $c_1 \supset c_2$ and $(g_1 \supseteq (c_1\backslash c_2) \cup g_2)$.*

Proof.

1. Soundness: if $c_1 \supset c_2$ and $g_1 \supseteq (c_1\backslash c_2) \cup g_2$, then r_1 is a NonReliable exact rule.
   According to the definition of Min-max exact basis, both r_1 and r_2 are Min-max exact rules. Since $g_1 \supseteq (c_1\backslash c_2) \cup g_2$, according to the definition of Reliable exact basis, r_1 must not be a Reliable exact rule, i.e., $r_1 \in NonReliableExact$.
2. Completeness: if r_1 is a NonReliable exact rule, then $c_1 \supset c_2$ and $g_1 \supseteq (c_1\backslash c_2) \cup g_2$.
   The proof follows the definition of $NonReliableExact$.

□

According to Theorem 3, for $r_2 : g_2 \Rightarrow c_2\backslash g_2$ where c_2 is a closed itemset and g_2 is a generator (i.e., r_2 is a rule in $MinMaxExact$), if we can find a super set $c_1 \supset c_2$, and $(g_1 \supseteq (c_1\backslash c_2) \cup g_2)$, then we can deduce: $r_1 : g_1 \Rightarrow c_1\backslash g_1$ is a $NonReliable$ basis rule. This means, from a rule in $MinMaxExact$, we can deduce a $NonReliable$ basis rule. Algorithm 3 given below takes $ReliableExact$ as the initial value for $MinMaxExact$ and generates all $NonReliable$ basis rules from $ReliableExact$ so that $MinMaxExact$

will be completed progressively during the course. Theorem 3 ensures that we can deduce all $NonReliable$ basis rules. On the completion of executing Algorithm 3, $MinMaxExact$ will contains all $ReliableExact$ basis rules and all $NonReliable$ basis rules as well.

For Algorithm 3, initially, the variable $MinMaxExact$ is assigned with $ReliableExact$. Steps 3-7 generate exact association rules from a basis rule in current $MinMaxExact$. Steps 8 to 13 deduce possible $NonReliable$ basis rules and add them into the current $MinMaxExact$.

Algorithm 3. AllExactFromReliable*(ReliableExact)*

Input: *ReliableExact: reliable exact basis*

Output: *AllExact: A set of all exact association rules*

1. $AllExact := \emptyset$, $MinMaxExact := ReliableExact$
2. for each rule $(r_1 : a_1 \Rightarrow c_1, r_1.supp) \in MinMaxExact$
3. *for each subset* $c_2 \subset c_1$
4. $AllExact := AllExact \cup \{(r_2 : a_1 \Rightarrow c_2, r_1.supp)\}$
5. *if* $(r_3 : a_1 \cup c_2 \Rightarrow c_1 \backslash c_2, r_1.supp) \notin AllExact$
6. *then* $AllExact := AllExact \cup \{r_3\}$
7. *end*
8. *for each super set* $c_3 \supset (c_1 \cup a_1)$ *and*
c_3 *is a closed itemset*
9. *for each* $a_3 \in G_{c_3}$
10. *if* $a_3 \supseteq ((c_3 \backslash (c_1 \cup a_1)) \cup a_1)$
11. *then* $MinMaxExact := MinMaxExact \cup \{a_3 \Rightarrow (c_3 \backslash a_3)\}$
12. *end*
13. *end*
14.end
15.return AllExact

4.2.2. Deriving Approximate Association Rules

Similar to the Min-max exact basis, the Reliable approximate basis is a subset of the Min-max approximate basis. Based on the definitions 6 and 8, the Min-max approximate basis can be described as:

$MinMaxApprox = \{g \Rightarrow (c \backslash g) | c \in C, g \in G, \gamma \circ \tau(g) \subset c\}$
$= \{g \Rightarrow (c \backslash g) | c \in C, g \in G_c, \neg(g \supseteq ((c \backslash c') \cup g'))$ or
$conf(g \Rightarrow c \backslash g) > conf(g' \Rightarrow c' \backslash g') for\ all\ c' \in C, g' \in G, \gamma \circ \tau(g) \subset c\} \cup$

$\{g \Rightarrow (c \backslash g) | c \in C, g \in G_c, g \supseteq ((c \backslash c') \cup g')$ and $conf(g \Rightarrow c \backslash g) \leq conf(g' \Rightarrow c' \backslash g')$
$for\ some\ c' \in C, g' \in G, \gamma \circ \tau(g') \subset c'\}$
$= ReliableApprox \cup NonReliableApprox$

Where $NonReliableApprox = \{g \Rightarrow (c \backslash g) | c \in C, g \in G_c, g \supseteq ((c \backslash c') \cup g')$ and $conf(g \Rightarrow c \backslash g) \leq conf(g' \Rightarrow c' \backslash g')$ $for\ some\ c' \in C, g' \in G, \gamma \circ \tau(g') \subset c'\}$

Similar to deriving exact rules, given $ReliableApprox$, if we can deduce $NonReliableApprox$ from $ReliableApprox$, then we would be able to deduce all approximate association rules. The following theorem shows that, for $r_2 : g_2 \Rightarrow c_2 \backslash g_2$, $c_2 \in C$ and $g_2 \in G$ (i.e., r_2 is a rule in $MinMaxApprox$), if for some $c_1 \in C$ and some $g_1 \in G$, $(g_1 \supseteq (c_1 \backslash c_2) \cup g_2)$ and $conf(r_1) \leq conf(r_2)$ are true, then we can deduce: $r_1 : g_1 \Rightarrow c_1 \backslash g_1$ is a rule in $NonReliableApprox$. This means that, from a rule in $MinMaxApprox$, we may deduce a $NonReliableApprox$ rule.

Theorem 4. *Let C be the set of frequent closed itemsets and G be the set of minimal generators. For rules $r_1 : g_1 \Rightarrow c_1 \backslash g_1$ and $r_2 : g_2 \Rightarrow c_2 \backslash g_2$ where c_1, $c_2 \in C$, g_1, $g_2 \in G$, $\gamma \circ \tau(g_1) \subset c_1$, and $\gamma \circ \tau(g_2) \subset c_2$. r_1 is a NonReliable approximate rule iff $(g_1 \supseteq (c_1 \backslash c_2) \cup g_2)$ and $conf(r_1) \leq conf(r_2)$.*

Proof.

1. Soundness: if $(g_1 \supseteq (c_1 \backslash c_2) \cup g_2)$ and $conf(r_1) \leq conf(r_2)$, then r_1 is a NonReliable approximate rule.
   According to the definition of Min-max approximate basis, both r_1 and r_2 are Min-max approximate rules. Since $g_1 \supseteq (c_1 \backslash c_2) \cup g_2$ and $conf(r_1) \leq conf(r_2)$, according to the definition of Reliable approx basis, r_1 must not be a reliable approximate rule. Therefore, $r_1 \in NonReliableApprox$.
2. Completeness: if r_1 is a NonReliable approximate rule, then $(g_1 \supseteq (c_1 \backslash c_2) \cup g_2)$ and $conf(r_1) \leq conf(r_2)$.
   The proof follows the definition of $NonReliableApprox$.

□

We designed the following algorithm $AllApproxFromReliable$ to derive all approximate rules from the Reliable Approximate basis. The algorithm $AllApproxFromReliable$ takes $ReliableApprox$ as the initial value for

$MinMaxApprox$. Steps 4-8 generate approximate rules from an approximate basis rule in current $MinMaxApprox$. Steps 9 to 14 deduce $NonReliableApprox$ basis rules and add them into the current $MinMaxApprox$. Therefore, during the process of deriving approximate rules, we generate all $NonReliableApprox$ rules so that $MinMaxApprox$ will be completed progressively during the course. Theorem 4 ensures that we can deduce all $NonReliableApprox$ basis rules. On the completion of executing Algorithm 4, $MinMaxApprox$ will contain all $ReliableApprox$ basis rules and also all $NonReliableApprox$ basis rules. Steps 17 to 21 in Algorithm 4 derive all approximate rules from these basis rules, which perform the same task as the steps 11 to 17 in the approximate reconstruction algorithm proposed in [PTB+05].

Algorithm 4. AllApproxFromReliable*(ReliableApprox)*

Input: *ReliableApprox: reliable approximate basis*

Output: *AllApprox: A set of all approximate association rules*

1. $AllExact := \emptyset$, $MinMaxApprox := ReliableApprox$
2. for $i = 2$ *to maximum size of closed itemsets*
3. for rule $(r_1 : a_1 \Rightarrow c_1, r_1.supp, r_1.conf) \in MinMaxApprox$ *and* $|c_1| = i$
4. for subset $c_2 \subset c_1$
5. if $(r_2 : a_1 \Rightarrow c_2, r_2.supp, r_2.conf) \notin AllApprox$
6. and $r_2.conf \neq 1$ *// *r_2 *is not an exact rule*
7. then $AllApprox := AllApprox \cup \{(r_2 : a_1 \Rightarrow c_2, r_1.supp, r_1.conf)\}$
8. end-for
9. for each closed itemset c_3
10. for generator a *such that* $a \supseteq a_1$ *and* $a.closure \subset c_3$
11. if $a \supseteq ((c_3 \backslash (c_1 \cup a_1)) \cup a_1)$ *and* $r_1.conf \geq \frac{c_3.supp}{a.supp}$
12. then $MinMaxApprox := MinMaxApprox \cup \{(a \Rightarrow (c_3 \backslash a), c_3.supp, \frac{c_3.supp}{a.supp})\}$
13. end-for
14. end-for
15. end-for
16. end-for
17. for rule $(r_1 : a_1 \Rightarrow c_1, r_1.supp, r_1.conf) \in AllApprox$
18. for each subset $c_3 \subseteq c_2$ *where* $c_2 = a_1.closure \backslash a_1$,

$\frac{(a_1.closure).supp}{a_1.supp} = 1$

19. $AllApprox := AllApprox \cup \{(a_1 \cup c_3 \Rightarrow c_1 \backslash c_3, r_1.supp, r_1.conf)\}$

20. end-for

21. end-for

22.return AllExact

In Algorithm 4, a rule is denoted as a 3-tuple (r, support, confidence), $r.supp$, $r.conf$, and $r.closure$ represent the support, confidence, and closure of r, respectively.

5. Experiments and Evaluation

We have conducted experiments to evaluate the effectiveness of the proposed Reliable bases. In this section, we first discuss the experimental hypotheses, then separately report on the experiments used to test the two hypotheses including descriptions of the datasets used in the experiments, evaluation metrics, experimental results and discussion.

5.1. Hypotheses

As discussed in the Introduction, three factors should be taken into consideration to evaluate the effectiveness of non-redundant rule mining algorithms. Firstly, the algorithm should be able to remove as much redundancy as possible. The smaller the size of the extracted rule set, the easier and more effective it is to use and maintain the rules. Secondly, all association rules should be able to be derived from the extracted non-redundant rules. This indicates that the extracted non-redundant rules capture all the information contained in the whole rule set. Instead of using the whole rule set, only using the non-redundant rules will not result in any loss of information for the application. Thirdly, the extracted non-redundant rules should retain the same inference capacity. That means, any problems that can be solved by using the whole rules can be solved by using the non-redundant rules. In Section 4, we have proved theoretically that the Reliable bases are more concise than the Min-Max bases and thus should contain a lesser number of basis rules. We have also proved that, from the Reliable bases, we can deduce all association rules. In this section, we will experimentally evaluate the Reliable bases in terms of the following two hypotheses which cover the three factors mentioned above.

- Hypothesis 1: The number of non-redundant rules generated by the Reliable bases is not larger than the number of non-redundant rules generated by the Min-Max bases for both exact rules and approximate rules. Both the Reliable bases and the Min-Max bases can deduce all association rules from the bases for both exact rules and approximate rules.

- Hypothesis 2: The non-redundant rules generated by the Reliable bases can provide a similar or even a higher capability to solve problems as compared to the non-redundant rules generated by the Min-Max bases.

5.2. Experiments for Testing Hypothesis 1

5.2.1. Datasets

The datasets used in the experiment for testing hypothesis 1 were obtained from UCI KDD Machine Learning Repository (http://kdd.ics.uci.edu/). The Mushroom dataset contains 8,124 transactions each of which describes the characteristics of one mushroom object. Originally each mushroom object has 23 attributes some of which are multiple value attributes. After converting the multiple value attributes to binary ones, the number of attributes of each object becomes 126. The other datasets include the Connect-4, Chess datasets which were derived from their respective game steps, the Breast Cancer dataset which was obtained from the University of Wisconsin Hospitals, the Annealing dataset containing annealing instances, and the Flare2 dataset containing solar flare instances each of which represents captured features for one active region on the sun. Table 5 shows some characteristics of the 6 datasets. Some of these datasets are very dense such as Connect-4 and Chess. They produce large numbers of frequent itemsets and thus a huge number of association rules even for very high values of support. Redundancy elimination is particularly important to these dense datasets.

5.2.2. Evaluation Metrics

Hypothesis 1 is concerned with the size of the extracted rule sets. The evaluation metric is straightforward. We simply check the number of rules generated by the Min-Max bases and the Reliable bases to compare their effectiveness in generating non-redundant rules. In order to measure the improvement achieved via use of the Reliable bases, we designed the following metrics to measure the volume of the rules reduced by using the Min-Max bases or the volume reduced by using the Reliable bases. Let N_{Total} be

Table 5: Dataset characteristics

Datasets	#transactions	#original attributes	#attributes after conversion
Mushroom	8,124	23	126
Connect-4	67,557	43	129
Chess	3,196	36	108
Breast Cancer	699	10	91
Annealing	898	38	276
Flare2	1,066	13	50

Table 6: Number of exact rules and reduction (Mushroom, minconf=0.5)

Minsupp	Total exact rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.3	2,142	453, 79%	117, 95%	74%
0.4	493	145, 71%	51, 90%	65%
0.5	161	44, 73%	18, 89%	60%
0.6	46	20, 57%	12, 74%	40%
0.7	27	12, 56%	6, 78%	50%
Average		67%	85%	58%

the number of total rules (exact rules or approximate rules), N_{MM} be the number of Min-Max basis rules (exact rules or approximate rules), and N_{RE} be the number of Reliable basis rules (exact rules or approximate rules), three reduction ratios are defined as below:

- Reduction ratio achieved by Min-Max against the entire rule set: $R^{MM}_{Total} = (N_{Total} - N_{MM})/N_{Total}$
- Reduction ratio achieved by Reliable against the entire rule set: $R^{RE}_{Total} = (N_{Total} - N_{RE})/N_{Total}$
- Reduction ratio achieved by Reliable against Min-Max: $R^{RE}_{MM} = (N_{MM} - N_{RE})/N_{MM}$

5.2.3. Results and Discussion

The experiment results are given in Table 6 to Table 17. For all tests, the *minconf* was set to 0.5. Table 6 to Table 11 present the test results

Table 7: Number of exact rules and reduction (Connect-4, minconf=0.5)

Minsupp	Total exact rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.94	4,862	1,110, 77%	50, 99%	95%
0.95	2,096	684, 67%	32, 98%	95%
0.96	746	354, 53%	15, 98%	96%
0.97	245	161,34%	7, 97%	96%
Average		58%	98%	96%

Table 8: Number of exact rules and reduction (Chess, minconf=0.5)

Minsupp	Total exact rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.90	132	116, 12%	3, 98%	97%
0.91	59	59, 0%	2, 97%	97%
0.93	32	32, 0%	2, 94%	94%
0.95	4	4, 0%	1, 75%	75%
Average		3%	91%	91%

of the exact bases, and Table 12 to Table 17 present the test results of the approximate bases. In this experiment, firstly we confirmed that from both the MinMax bases and the Reliable bases we can deduce all exact rules and all approximate rules as indicated in the second column of tables 6 to 17. For example, when $Minsupp$ is 0.3, both MinMax and Reliable bases produce 2,142 exact rules and 21,377 approximate rules for the Mushroom dataset as showed in Table 6 and Table 12, respectively. Secondly, we tested the reduction ratio between the size of the MinMax bases and the size of Reliable bases for different $Minsupp$ settings.

We surprisingly found that the reduction ratios achieved by the Reliable bases against the Min-Max bases are very high. As indicated by the average reduction ratios displayed in the bottom row of tables 6 to 12, for exact rules, the highest average reduction ratio is 96% for the Connect-4 dataset. Even for the lowest average reduction ratio which is 58% for the Mushroom dataset, more than half of the exact rules generated by the Min Max base are considered redundant and therefore not generated by the Reliable base. The

Table 9: Number of exact rules and reduction (Breast cancer, minconf=0.5)

Minsupp	Total exact rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.2	79	79, 0%	10, 87%	87 %
0.3	74	74, 0%	9, 88%	88%
0.4	25	25, 0%	7, 72%	72%
Average		0%	82%	82%

Table 10: Number of exact rules and reduction (Annealing, minconf=0.5)

Minsupp	Total exact rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.3	650	194, 70%	44, 93%	77%
0.4	265	89, 66%	23, 91%	74%
0.5	104	44, 58%	7, 93%	84%
0.6	44	28, 36%	6, 86%	79%
Average		58%	91%	79%

highest reduction ratio was obtained for the Chess dataset for $Minsupp$=0.9. In this case, the Min-Max exact basis generates 116 exact rules as shown in Table 8, while the Reliable exact basis generates only 3 exact rules. 113 out of the 116 rules are considered redundant by the Reliable basis and the reduction ratio is 97% which is a very significant reduction. For the approximate bases, the reduction is also considerably high. The highest average reduction ratio is 77% for the Connect-4 dataset. As an example, we can see from Table 13, when $Minsupp$ was set to 0.94, the Min-Max approximate basis generates 49,407 approximate basis rules, while the Reliable approximate basis generates 10,220 basis rules with a reduction ratio of 79%.

After carefully checking the rules in each of the bases, we found that there indeed exists a great amount of redundancy in the Min-Max basis for each of the tests we conducted. For example, in the case of $Minsupp = 0.5$ for the Annealing dataset, for the rule *carbon-00* $\Rightarrow$ *product-type-C* in the Reliable exact basis, we found that the following 13 rules in the Min-Max exact basis are redundant to *carbon-00* $\Rightarrow$ *product-type-C*:

Table 11: Number of exact rules and reduction (Flare2, minconf=0.5)

Minsupp	Total exact rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.3	957	241, 75%	41, 96%	83%
0.4	364	154, 58%	47, 87%	69%
0.5	383	107, 72%	20, 95%	81%
0.6	230	90, 61%	11, 95%	88%
Average		66%	93%	80%

carbon-00,hardness-00 ⇒ *product-type-C*
carbon-00,strength-000 ⇒ *product-type-C*
carbon-00,bore-0000 ⇒ *product-type-C*
carbon-00,class-3 ⇒ *product-type-C*
carbon-00,hardness-00,strength-000 ⇒ *product-type-C*
carbon-00,hardness-00,bore-0000 ⇒ *product-type-C*
carbon-00,hardness-00,class-3 ⇒ *product-type-C*
carbon-00,strength-000,bore-0000 ⇒ *product-type-C*
carbon-00,strength-000,class-3 ⇒ *product-type-C*
carbon-00,bore-0000,class-3 ⇒ *product-type-C*
carbon-00,hardness-00,strength-000,bore-0000 ⇒ *product-type-C*
carbon-00,hardness-00,bore-0000,class-3 ⇒ *product-type-C*
carbon-00,strength-000,bore-0000,class-3 ⇒ *product-type-C*

Similarly for the approximate bases, for example, the following 9 rules in the $MinMax$ approximate basis are redundant to the reliable rule *steel-A* ⇒ *product-type-C,strength-000* (0.4844, 0.9886) in the Reliable approximate basis:

steel-A,carbon-00 ⇒ *product-type-C,strength-000, (0.47327,0.9884)*
steel-A,hardness-00 ⇒ *product-type-C,strength-000, (0.30512,0.9821)*
steel-A,bore-0000 ⇒ *product-type-C,strength-000, (0.4655,0.9882)*
steel-A,class-3 ⇒ *product-type-C,strength-000, (0.3853,0.9858)*
steel-A,carbon-00,bore-0000 ⇒ *product-type-C,strength-000, (0.4543,0.9879)*
steel-A,carbon-00,class-3 ⇒ *product-type-C,strength-000, (0.3775,0.9854)*
steel-A,hardness-00,bore-0000 ⇒ *product-type-C,strength-000, (0.3040,0.9820)*
steel-A,bore-0000,class-3 ⇒ *product-type-C,strength-000, (0.3731,0.9853)*

steel-A,carbon-00,bore-0000,class-3 ⇒ product-type-C,strength-000, (0.3653,0.9850)

The 13 exact rules in the Min-Max exact basis have the same consequent but a larger antecedent than those of the rule *carbon-00 ⇒ product-type-C*, their support values are different, but they have exactly the same confidence and the same CF value. In real world problem solving, if we know that *carbon-00* is true, by applying the rule *carbon-00 ⇒ product-type-C*, we can conclude that *product-type-C* is true. We don't have to know *hardness-00*, *strength-000*, or *bore-0000*, etc. in order to reach this conclusion. That means, all the 13 rules are useless if we have the rule *carbon-00 ⇒ product-type-C* at hand. Similarly, for the 9 approximate rules in the Min-Max approximate basis, they have the same consequent but a larger antecedent than that of the reliable rule *steel-A ⇒ product-type-C,strength-000*. Both the support and confidence values, as indicated as (support, confidence) at the end of each rule, of these 9 rules are smaller than those of the reliable rule. Therefore, according to Theory 1, their CF value won't be greater than that of the reliable rule. Therefore, if we know that *steel-A* is true, by applying the rule *steel-A ⇒ product-type-C,strength-000*, we can conclude that *product-type-C,strength-000* is true. We don't have to know *hardness-00*, *class-3*, or *bore-0000*, etc. in order to reach this consequence. That means, all the 9 rules are useless or redundant if we have the rule *steel-A ⇒ product-type-C,strength-000* at hand. By eliminating these redundant rules, the size of the *Reliable* bases is much smaller than that of the Min-Max bases, but the capacity of solving problems remains the same. This reduction provides a great potential to improve the effectiveness of using the extracted association rules.

The shortcoming of the proposed method is its efficiency. The complexity of the proposed Reliable exact and approximate algorithms are $O(n^2)$ and $O((mn)^2)$, respectively, while the Min-Max exact and approximate algorithms are $O(n)$ and $O(mn)$, respectively, where n is the number of generators and m is the number of closed itemsets. For large datasets, the proposed algorithms may have efficiency problems to generate the non-redundant Reliable basis rules. However, since the number of closed itemsets and the number of generators are usually much smaller than that of frequent itemsets, generating the non-redundant Reliable basis rules will still be more efficient than generating the entire rules using frequent itemsets. The proposed algorithms have the potential to be implemented in a parallel way. For both the reliable exact and approximate bases, the non-redundant association rules

Table 12: Number of approximate rules and reduction (Mushroom, minconf=0.5)

Minsupp	Total approx rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.3	21,377	2,634, 88%	1,970, 91%	25%
0.4	2,528	465, 82%	361, 86%	22%
0.5	835	175, 79%	135, 84%	23%
0.6	228	59, 74%	52, 77%	12%
0.7	161	39, 76%	34, 79%	13%
Average		80%	83%	19%

Table 13: Number of approximate rules and reduction (Connect-4, minconf=0.5)

Minsupp	Total approx rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.94	199,560	49,407, 75%	10,220, 95%	79%
0.95	77,206	24,794, 68%	5,245, 93%	79%
0.96	26,856	11,452, 57%	2,538, 91%	78%
0.97	7,895	4,439, 44%	1,214, 85%	73%
Average		61%	91%	77%

are generated from and only from the closed itemsets and their corresponding minimal generators, and the process of generating rules from one closed itemset is independent from the process of generating other rules using other closed itemsets. Moreover, during the whole process, the closed itemsets and minimal generators remain unchanged. Therefore, a possible approach could be to divide the set of closed itemsets into several subsets and conduct the rule mining on the subsets in parallel. The efficiency issue will be addressed in our future work and detailed parallel algorithms will be developed in particular.

5.3. Experiments for Testing Hypothesis 2

Recommender systems have been widely used in many e-commerce sites to help users find products or items of interest [SKKR00]. The most popularly used technique is the collaborative filtering method [SKR01] that makes rec-

Table 14: Number of approximate rules and reduction (Chess, minconf=0.5)

Minsupp	Total approx rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.90	1,0614	8,371, 21%	2,483, 77%	70%
0.91	5,785	5,050 , 13%	1,571, 73%	69%
0.93	2,338	1,948, 17%	688, 71%	65%
0.95	468	459, 2%	196, 58%	57%
Average		13%	70%	65%

Table 15: Number of approximate rules and reduction (Breast cancer, minconf=0.5)

Minsupp	Total approx rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.2	5,917	3,661, 38%	2,592, 56%	26%
0.3	5,134	3,269, 36%	2,391, 53%	27%
0.4	1,207	966, 20%	851, 29%	12%
Average		31%	46%	22%

ommendations based on users' previous ratings to products (also called user profiles). Recommender systems usually work effectively when user ratings are extensive and the applicable dataset has a high information density. One of the shortcomings of the collaborative-filtering recommendation approach is that it must be initialized with a large amount of user's rating data in order to make meaningful recommendations. When there is insufficient rating data, e.g., a user has very few ratings in their profile, recommender systems may fail to provide recommendations that interest the user. In this section, we report the experimental results that show that applying association rules to product recommendation can improve the quality of recommendations. Especially, we show that using the Reliable rules results in the same or better performance as compared to using the Min-max rules even the number of the Reliable rules is less than that of the Min-Max rules.

Table 16: Number of approximate rules (Annealing, minconf=0.5)

Minsupp	Total approx rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.3	5,052	865, 83%	554, 89%	36%
0.4	1,835	435, 76%	296, 84%	32%
0.5	1,186	300, 75%	218, 82%	27%
0.6	416	137, 67%	102, 75%	26%
Average		75%	83%	30%

Table 17: Number of approximate rules and reduction (Flare2, minconf=0.5)

Minsupp	Total approx rules (N_{Total})	MinMax (N_{MM},R^{MM}_{Total})	Reliable (N_{RE},R^{RE}_{Total})	R^{RE}_{MM}
0.3	7,604	1216, 84%	710,91%	42%
0.4	2,420	644, 73%	479, 81%	30%
0.5	5,599	1081, 81%	730, 87%	34%
0.6	5,368	1203, 78%	687, 87%	43%
Average		79%	87%	37%

5.3.1. User Profile Expansion Using Association Rules

For collaborative filtering recommender systems, a user's profile consists of the user's ratings of items. Very often the number of items is huge and only a very small number of items were rated by each user. Ziegler et al. have proposed a technique to convert users' item ratings to item categories ratings [ZLST04]. The resulting user profiles are more dense because the number of categories is usually much smaller than that of items. In this paper, by using Ziegler's technique we construct a dataset consisting of users' item category ratings. Even though the new dataset is more dense than the original dataset, there still exists many users who have a short list of categories. In order to improve the quality of recommendations being made to users with short profiles, we propose to use association rules to expand user profiles. We firstly construct a transactional dataset, each of the transactions in the dataset consists of the categories that interest a user. Secondly we mine the transactional dataset for association rules between categories. Each rule represents the association between two sets of categories. These

rules allow us to discover categories that may also be of interest to users. Finally, we expand the user profiles with the association rules. For each user profile which provides a list of categories that the user is interested, we generate all combinations from the categories. A rule whose antecedent is matched with one of the combinations will be used to expand the user's profile by adding the rule's consequent, which is a set of categories, into the profile. Our experiments show that the expanded user profiles have the potential to improve recommendation quality over profiles that have not been expanded and, most importantly, the non-redundant rules discovered using our proposed methods can achieve the same as or even better performance than the rules discovered using the Min-Max methods.

5.3.2. Datasets

For this investigation, we use the BookCrossing dataset obtained from http://www.informatik.uni-freiburg.de/cziegler/BX which contains 278,858 users, 271,379 books and about 1,149,780 ratings given to those books by the users. For the purpose of evaluation, each user profile (i.e., a set of ratings) was split into two parts with a ratio of 50-50. 50% of the ratings form a training dataset and the remaining 50% forms a test dataset. The training dataset is converted into a dataset containing user category ratings using Ziegler's conversion technique [ZLST04]. The book taxonomy data for converting users' book ratings to category ratings is obtained from Amazon.com web site. Not every book in the BookCrossing dataset is available in Amazon.com, we were only able to extract taxonomy data for 270,868 books. After the conversion, a transactional dataset is constructed that contains 85,415 transactions (i.e., distinct users) involving 270,868 unique books and 10,662 taxonomy categories. The average number of categories in a user profile is 27.08 and the highest number of categories in a user profile is 3,173. This set of user profiles will serve as our baseline dataset and is also the dataset that will be expanded using the derived association rules. The test data set will be used to evaluate the quality of recommendations.

5.3.3. Evaluation Metrics

The recommender used in this experiment is the Taxonomy-driven Product Recommender proposed in [ZLST04] and implemented using the Taste framework (http://taste.sourceforge.net/). The goal of this experiment is to evaluate the capacity of the rules generated by the Min-Max and the Reliable algorithms through comparing the quality of the recommendations generated

based on various user profiles including the baseline dataset and the expanded datasets obtained by using the Min-Max and the Reliable rules. For a user u_i, the recommender system will recommend a list of items, denoted as P_i, based on the user profiles. The recommendation list P_i will then be evaluated against the test dataset using the evaluation metrics precision, recall and F1-measure [HKTR04] which are defined below.

For a user u_i and an item t, let T_i be the set of items in the test dataset that are rated by the user, $rating(u_i, t)$ denote the rating that the user gave the item and $avg(u_i)$ denote the average of the user's ratings, we define the set of items that are preferred by the user as:

$$\widehat{T_i} = \{t | t \in T_i, rating(u_i, t) > avg(u_i)\}$$

The precision, recall, and F1 measure are calculated using the following equations:

$$Precision = \frac{|\widehat{T_i} \bigcap P_i|}{|P_i|}$$

$$Recall = \frac{|\widehat{T_i} \bigcap P_i|}{|\widehat{T_i}|}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

5.3.4. Results and Discussion

To test the hypothesis 2, we conducted a series of experiments to determine the improvement in the recommender system obtained by using different rule sets. The dataset is extremely sparse. In order to find frequent patterns and thus rules, we set the minimum support to 0.06. We mine the transactional dataset for association rules using the two non-redundant rule mining algorithms with minimum confident threshold being set to 0.5. Table 18 shows the number of rules found by the two algorithms, from which we can see that the entire rule set and the non-redundant rule set generated by the Min-Max algorithm are actually identical. This means that, for this sparse dataset, none of the rules discovered by the Min-Max algorithm are considered redundant based on the Min-Max redundancy definition, while our Reliable algorithm finds and removes 6% of the rules which are considered redundant based on Definition 5.

Table 18: Number of rules and reduction (BookCrossing, minconf=0.5, minSup=0.06)

Algorithm	Total rules	Non-redundant rules	R^{RE}_{MM}
MinMax	9,520	9,520	
Reliable	9,520	8,950	6%

In these experiments, the user profiles with 5 or less ratings are considered short user profiles. This yields a total of 15,912 short user profiles which are expanded using the discovered association rules. When expanding a user profile, it is possible that a large number of rules are matched with the user's profile. In the experiments, we select the top rules based on the confidence of the rules to expand user profiles. Table 19 to Table 21 show the precision, recall and F1 measure of the recommendations produced based on user profiles including the baseline profile and the profiles expanded using the top 1, 2, 3, 4 or 5 rules selected based on the confidence of the rules, where MM and RE stand for the Min-Max and the Reliable methods, respectively. As shown in the tables, the precision, recall, and F1 measure of the recommendations produced by using the expanded datasets are improved compared to those produced by using the baseline dataset, which means that the association rules are useful in improving the quality of the user profiles. The non-redundant rules can be used in place of a larger rule set that contains redundancy without degrading performance. The results also indicate that the performance of the recommender improves as more rules are used in expanding the user profiles. For instance, by using the rules generated by the Reliable algorithm to expand the user profiles, the precision of the recommender is improved over the baseline by approximately 4.85% for using 1 rule to 32.31% for using the top 5 rules. Most importantly, the results clearly show that the performance of the recommender system is very similar for all the three measures when using different user profiles expanded by the two algorithms (MM and RE). Actually, for the BookCrossing dataset, the recommender achieved better results when using the Reliable rule set than using the Min-Max rule set for all the three measures even though the size of the Reliable rule set is 6% smaller than that of the Min-Max rule set. This reinforces the hypothesis that a smaller non-redundant rule set can be used instead of a larger rule set that contains redundancy. Therefore, we believe that the experiment results strongly support the hypothesis 2 that the non-redundant rules generated by the Reliable bases can provide simi-

Table 19: Precision of recommendations based on various user profiles

User Profiles	Precision		Improvement over Baseline		Improvement
	MM	RE	MM	RE	RE over MM
Baseline	0.00619				
Expanded (top 1 rule)	0.00649	0.00649	4.85%	4.85%	0%
Expanded (top 2 rules)	0.00714	0.00714	15.35%	15.35%	0%
Expanded (top 3 rules)	0.00732	0.00734	18.26%	18.58%	1.77%
Expanded (top 4 rules)	0.00792	0.00798	27.95%	28.92%	3.47%
Expanded (top 5 rules)	0.00815	0.00819	31.66%	32.31%	2.04%

Table 20: Recall of recommendations based on various user profiles

User Profiles	Recall		Improvement to Baseline		Improvement
	MM	RE	MM	RE	RE over MM
Baseline	0.0571				
Expanded (top 1 rule)	0.0596	0.00596	4.38%	4.38%	0%
Expanded (top 2 rules)	0.0655	0.0655	14.71%	14.71%	0%
Expanded (top 3 rules)	0.0673	0.0674	17.86%	18.04%	0.98%
Expanded (top 4 rules)	0.073	0.0736	27.85%	28.90%	3.77%
Expanded (top 5 rules)	0.0749	0.0754	31.17%	32.05%	2.81%

lar or even higher capability to solve problems as the non-redundant rules generated by the Min-Max bases. It is also supports the theory behind the non-redundant rule mining algorithms, that the redundant rules removed by these algorithms do not contain new information that can not be captured by the non-redundant rules.

6. Related Work

One approach to address the quality of association rules is to apply constraints to generate only those association rules that are interesting to users. [NLHP98] and [SVA97] proposed some algorithms that incorporate item constraints to the process of generating frequent itemsets. Some work has also

Table 21: F1 measure of recommendations based on various user profiles

User Profiles	F1		Improvement to Baseline		Improvement
	MM	RE	MM	RE	RE over MM
Baseline	0.0112				
Expanded (top 1 rule)	0.0117	0.0117	4.51%	4.51%	0%
Expanded (top 2 rules)	0.0129	0.0129	14.97%	14.97%	0%
Expanded (top 3 rules)	0.01320	0.01324	17.89%	18.20%	1.72%
Expanded (top 4 rules)	0.0143	0.0144	27.59%	28.56 %	3.53%
Expanded (top 5 rules)	0.0147	0.0148	31.25%	31.92%	2.13%

been done on measuring association rules with interestingness parameters [BMUT97]. These approaches focus on pruning the association rules to get more general or informative association rules based on interestingness parameters. The approach proposed in [BAG00] integrates various constraints into the mining process including consequent constraint and minimal improvement constraint. The consequent constraint is used to restrict rules with certain consequent specified by the user. The minimal improvement constraint is used to simplify the antecedents of rules based on items' contribution to the confidence and therefore prune association rules that have more specific antecedent but do not make more contribution to the confidence. Another approach is to use a taxonomy of items to extract generalized association rules [HF00], i.e., to generate rules between itemsets that belong to different abstract levels in the taxonomy, especially between high abstract levels, aiming at reducing the number of extracted rules. The approaches mentioned above aim to reduce the number of extracted rules and also improve the"usefulness" of the rules, but eliminating redundancy of rules is not a focus. The approaches proposed in [Zak04], [PTB$^+$05], [GMT05], and [CG07] focus on extracting non-redundant itemsets and association rules. Recently, the investigation of redundancy elimination has been extended to multi-level datasets [SXG08] and sequential datasets [LKW09] which are related but beyond the scope of this paper. In the rest of this section, we will focus on discussing the approaches proposed in [Zak04], [PTB$^+$05], [GMT05], [CG07], and some other relevant works.

[Zak04] and [PTB$^+$05] make use of the closure of the Galois connection [GW99] to extract non-redundant rules from frequent closed itemsets instead

of from frequent itemsets. The difference between the two approaches is the definition of redundancy. The approach proposed in [Zak04] extracts rules, called the most general rules, that have the shortest antecedent and shortest consequent in an equivalent class of rules with the same confidence and the same support. All other rules in the equivalent class are considered redundant to the extracted rules. The extracted rules (i.e., the most general rules) constitute a generating rule set from which all other rules can be derived. However, the generating set may not retain the same inference capacity as the entire rule set. For example, for the equivalent class of rules $TW \Rightarrow A$, $TW \Rightarrow AC$, and $CTW \Rightarrow A$ given in [Zak04], the most general rule in this class is $TW \Rightarrow A$, the other two rules are considered redundant to this rule and won't be included in the generating set. But by using the extracted rule $TW \Rightarrow A$ we can't derive the consequent that can be derived by using rule $TW \Rightarrow AC$ which is considered redundant. Therefore, using the extracted rules alone can't guarantee the same results derived by using the entire rule set. However, this problem won't occur when using the methods proposed in this paper and [PTB+05] as well. The non-redundant rule set generated by the methods proposed in this paper and [PTB+05] retain the same capacity as that of the entire rule set. While both [PTB+05] and this paper define the non-redundant rules are those which have minimal antecedents and maximal consequents in an equivalent class, our definition relaxed the requirement to redundancy that the redundant rules don't have to have the same support and confidence as their corresponding non-redundant rules. The relaxed requirement allows more rules to be considered redundant and thus eliminated. Most importantly, we proved that the elimination of such redundant rules does not reduce the belief of the extracted rules and the capacity of the extracted rules for solving problems is also not reduced.

The concept of non-derivable itemsets was first introduced in [CG02] and further studied in [CG07]. Based on the inclusion-exclusion principle [Knu97], Calders and Goethals [CG02] proposed a method to derive a lower and an upper bound on the support of an itemset from the supports of all its subsets. When these bounds are equal, the itemset is considered derivable. The itemsets whose lower bound and upper bound are different are called non-derivable itemsets from which the supports of all derivable itemsets can be derived and as such the non-derivable itemsets form a concise representation of all itemsets. Another important concept is the closed non-derivable itemset proposed by Muhonen and Toivonen in [MT06]. A closed non-derivable itemset is defined as the closure of a non-derivable itemset.

But unlike the closed frequent itemsets which must be frequent itemsets, the closed non-derivable itemsets are not necessarily non-derivable itemsets. However, it was proved in [MT06] that for a given dataset, the number of closed non-derivable itemsets is smaller than or equal to the number of the non-derivable itemsets and also smaller than or equal to the number of closed itemsets. Furthermore, an algorithm was also proposed in [MT06] to derive all frequent itemsets and their supports from the closed non-derivable itemsets. For concisely representing frequent itemsets, there is no doubt that the set of closed non-derivable itemsets as well as the set of non-derivable itemsets is a concise representation of all itemsets. Since the set of closed non-derivable itemsets is smaller than that of the closed itemsets and the non-derivable itemsets as well [MT06], it is a more concise representation than closed itemsets and non-derivable itemsets.

However, for generating concise representations of association rules which is the focus of this paper, no work has been found that generates concise representations of association rules based on the non-derivable itemsets or closed non-derivable itemsets except for the work done by Goethals et al. [GMT05] which presents methods for deriving the lower and upper bounds of confidence of an association rule from the supports and confidences of all its subrules. A rule is considered derivable if the lower bound and the upper bound of its confidence are equal. Because the confidence of a derivable rule can be derived given its subrules, it is considered redundant with respect to its subrules [GMT05]. According to [GMT05], the non-derivable rules can be generated using the non-derivable itemsets. One important feature of a concise representation of association rules is that the representation is a generating set from which all other rules can be derived. However, the set of non-derivable association rules can't be used as a generating set to derive all other rules. If the user wants other rules, traditional techniques have to be used to generate all rules by re-scanning the dataset. Another important difference between the non-derivable association rules and the Reliable basis rules proposed in this paper is that the set of non-derivable rules may not provide the same capacity as that of the entire rule set. Therefore, unlike the Reliable basis and the Min-Max basis, the set of non-derivable rules can't replace the entire rule set in solving problems. For example, for the simple dataset $\{abcd, cd, ab, abd, ac, acd\}$ with 6 transactions, d and ad are two non-derivable itemsets from which we can generate a non-derivable rule $d \Rightarrow a$ with confidence 0.67. But this rule is considered redundant by the Reliable method and the Min-Max method as well since these two methods

can generate a rule $d \Rightarrow ac$ which has the same confidence (i.e., 0.67) as that of $d \Rightarrow a$ but derives a longer consequent (i.e., ac instead of a). In contrast, for the non-derivable rule mining, $d \Rightarrow ac$ is considered redundant since its confidence can be derived from its subrules and therefore eliminated. We argue that the rule $d \Rightarrow ac$ is more useful than $d \Rightarrow a$ since it derives a longer consequent than that derived by $d \Rightarrow a$ with the same confidence. On the other hand, some non-derivable rules are actually redundant according to the Reliable and Min-Max redundancy definitions. Still using the same example above, b, bd, and abd are non-derivable itemsets from which we can generate two non-derivable rules $b \Rightarrow d$ and $b \Rightarrow ad$, both rules have the confidence 0.33. Obviously, rule $b \Rightarrow d$ concludes less information than that of rule $b \Rightarrow ad$ and therefore is considered redundant to $b \Rightarrow ad$ in terms of the Reliable or Min-Max redundancy definition. In fact, both rules are considered redundant because both the Reliable and the Min-Max algorithms can generate a non-redundant rule $b \Rightarrow acd$ with exactly the same confidence 0.33. The consequent produced by using rule $b \Rightarrow acd$ covers the consequent produced by using $b \Rightarrow d$ or $b \Rightarrow ad$. Therefore $b \Rightarrow d$ and $b \Rightarrow ad$ are actually useless if $b \Rightarrow acd$ is provided. However, $b \Rightarrow acd$ is considered a derivable rule and won't be included in the non-derivable rule set.

The certainty factors have been used to determine useful association rules in many works. For example, in [BBSV02, DSMB01], the authors pointed out the insufficiency of only using support and confidence to determine useful rules. They proposed to choose rules based on both certainty factors and supports. They conducted experiments on large medical datasets and have shown good results in practice [DSMB01]. In this paper we proposed to use the certainty factor as the criterion to measure the strength of the discovered association rules. The difference from previous works is that we use the certainty factor to verify that the redundancy elimination proposed in this paper will not damage the quality of the extracted rules. We proved that the elimination of the redundant rules defined in this paper will not reduce the certainty factor values of the extracted rules.

7. Conclusion

One challenging problem with association rule mining is the redundancy existing in the extracted association rules which greatly impacts the effective use of the extracted rules in solving real world problems. A satisfactory solution to the problem should be one that can maximally remove redundancy

but does not damage the inference capacity of and the belief in the extracted rules. Moreover, an appropriate criterion to define a boundary between redundancy and non-redundancy is desirable. Since the late 1990s, many efforts have been made to improve the quality of association rules by eliminating redundancy. The approaches proposed in [Zak04] and [PTB$^+$05] are successful approaches. Both approaches generate association rules from frequent closed itemsets instead of from frequent items. The approach proposed in [Zak04] generates the most general rules which have the shortest antecedent and shortest consequent in an equivalent class, while the approach proposed in [PTB$^+$05] generates the Min-max rules which have the shortest antecedent and longest consequent in an equivalent class. Both approaches can significantly remove redundancy. However, as we have pointed out that the most general rules generated in [Zak04] may not retain the same inference capacity as the entire rule set and the Min-max rules generated in [PTB$^+$05] still contain redundancy. In this paper, a concise representation of association rules called Reliable basis was presented which can ensure the removal of the maximal amount of redundancy without reducing the inference capacity of the remaining extracted rules. Moreover, we proposed to use the certainty factor as the criterion to measure the strength of the discovered association rules. Based on the certainty factor theory, we theoretically proved that the Reliable basis contains no redundancy and the strength of belief in the associations captured by the extracted Reliable basis rules are not less than that of the entire set of rules. We also experimentally demonstrated that the proposed Reliable basis retains the same inference capacity as the entire rule set. Furthermore, we theoretically proved and experimentally confirmed that the proposed Reliable basis is not only concise but also lossless because all association rules can be retrieved from the Reliable basis. The time complexity of the proposed algorithms to generate the Reliable basis rules is higher than that of generating the Mix-max basis rules. Developing parallel algorithms could be a way to improve the efficiency of generating the Reliable basis rules. This issue will be addressed in our future work.

Acknowledgment

The authors would like to thank the referees for their valuable comments and suggestions.

References

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on very large data bases*, pages 487–499, 1994.

[BAG00] R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4:217–240, 2000.

[Bay98] R. J. Bayardo. Efficiently mining long patterns from databases. In *Proceedings of the 1998 ACM SIGMOD Conference*, pages 85–93, 1998.

[BBR03] Jean-Francois Boulicaut, Artur Bykowski, and Christophe Rigotti. Free-sets: a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery*, 7:5–22, 2003.

[BBSV02] F. Berzal, I. Blanco, D. Sanchez, and M. A. Vila. Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*, 6:221–235, 2002.

[BL97] M. J. A. Berry and G. S. Linoff. *Data Mining Techniques for Marketing, Sales and Customer Support*. John Wiley and Sons, 1997.

[BMUT97] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD Conference*, pages 255–264, 1997.

[BR03] A. Bykowski and C. Rigotti. Dbc: A condensed representation of frequent patterns for efficient mining. *Information Systems*, 28:949–977, 2003.

[CCL05] Alain Casali, Rosine Cicchetti, and Lotfi Lakhal. Essential patterns: A perfect cover of frequent patterns. In *DaWaK*, pages 428–437, 2005.

[CG02] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. pages 74–85. Springer, 2002.

[CG07] Toon Calders and Bart Goethals. Non-derivable itemset mining. *Data Mining and Knowledge Discovery*, 14(1):171–206, 2007.

[CRB06] T. Calders, C. Rigotti, and J-F. Boulicaut. A survey on condensed representations for frequent sets. In J-F. Boulicaut, L. de Raedt, and H. Mannila, editors, *Constraint-Based Mining*, volume 3848 of *LNCS*, pages 64–80. Springer, 2006.

[DD09] Dietmar H. Dorr and Anne M. Denton. Establishing relationships among patterns in stock market data. *Data and Knowledge Engineering*, 68(3):318–337, 2009.

[DSMB01] M. Delgado, D. Sanchez, and M-J. Martin-Bautista. Mining association rules with improved semantics in medical databases. *Aritificial Intelligence in Medicine*, 21:241–245, 2001.

[GMT05] B. Goethals, J. Muhonen, and H. Toivonen. Mining non-derivable association rules. In *Proceedings of the SIAM International Conference on Data Mining*, pages 239–249, 2005.

[GW99] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, 1999.

[HF00] J. Han and Y. Fu. Mining multiple-level association rules in large databases. *IEEE Transactions on Knowledge and Data Engineering*, 11:798–804, 5 2000.

[HKTR04] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, January 2004.

[HLN99] Jiawei Han, Laks V. S. Lakshmanan, and Raymond T. Ng. Constraint-based multidimensional data mining. *IEEE Computer*, 32(8):46–50, 1999.

[HP00] J. Han and J. Pei. Mining frequent patterns by pattern-growth: methodology and implications. *ACM SIGKDD Explorations Newsletter*, 2(2):14–20, 2000.

[Knu97] D.E. Knuth. *Fundamental Algorithms*. Addison-Wesley, Massachusetts, 1997.

[KRG04] M. Kryszkiewicz, H. Rybinski, and M. Gajek. Dataless transitions between concise representations of frequent patterns. *Journal of Intelligent Information Systems*, 22(1):41–70, 2004.

[LKW09] David Lo, Siau-Cheng Khoo, and Limsoon Wong. Non-redundant sequential rules - theory and algorithm. *Information Systems*, 34:438–453, 2009.

[MSMG04] Sally McClean, Bryan Scotney, Philip Morrow, and Kieran Greer. Knowledge discovery by probabilistic clustering of distributed databases. *Data and Knowledge Engineering*, 54(2):189–210, 2004.

[MT06] Juho Muhonen and Hannu Toivonen. Closed non-derivable itemsets. In *Proceedings of PAKDD'06*, pages 601–608, 2006.

[NLHP98] R. T. Ng, V.S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning otimizations of constrained association rules. In *Proceedings of the SIGMOD conference*, pages 13–24, 1998.

[PBTL99a] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th ICDT conference*, pages 398–416, 1999.

[PBTL99b] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, 1999.

[PTB+05] N. Pasquier, R. Taouil, Y. Bastide, G. Stumme, and L. Lakhal. Generating a condensed representation for association rules. *Journal of Intelligent Information Systems*, 24(1):29–60, 2005.

[SB75] E. H. Shortliffe and B. G. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3/4):351–379, 1975.

[SKKR00] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In *Proceedings of*

the 2nd ACM conference on Electronic commerce, pages 158–16, 2000.

[SKR01] J. Schafer, J. Konstan, and J. Riedi. E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, 5(1-2):115–153, 2001.

[SVA97] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In *Proceedings of the KDD Conference*, pages 67–73, 1997.

[SXG08] Gavin Shaw, Yue Xu, and Shlomo Geva. Deriving non-redundant approximate association rules from hierarchical datasets. In *Proceeding of the 17th ACM conference on Information and knowledge management (CIKM'08)*, pages 1451–1452, 2008.

[Wil82] R Wille. *Restructuring lattices theory: An approach based on hierarchies of concepts*. I. Rival (editor), Ordered sets. Dordrecht-Boston, 1982.

[Zak00] M. J. Zaki. Generating non-redundant association rules. In *Proceedings of the KDD Conference*, pages 34–43, 2000.

[Zak04] M. J. Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9:223–248, 2004.

[ZLST04] C.-N. Ziegler, G. Lauser, and L. Schmidt-Thieme. Taxonomy-driven computation of product recommendations. In *International Conference on Information and Knowledge Management (CIKM'04)*, pages 406–415, 2004.