VIE_MGB: A Visual Interactive Exploration of Minimal Generic Basis of Assopciation Rules

Chiraz Latiri Cherif[1], Wissem Bellegha[1]
Sadok Ben Yahia[2] and Ghada Guesmi[2]

[1] École Supérieure de Commerce de Tunis,
Computer Sciences Department,
Campus Universitaire de La Manouba,
2010, Tunisia
chiraz.latiri@gnet.tn
wissembellegha@gmail.com

[2] Faculty of Sciences of Tunis, Computer Sciences Department
Research Team URPAH
Campus Universitaire EL Manar, 1060 Tunis, Tunisia
sadok.benyahia@fst.rnu.tn
ghada_gasmi@yahoo.fr

Abstract. Mining association rules is an important task, even though the number of rules discovered is often huge. A possible solution to this problem, is to use the Formal Concept Analysis (FCA) mathematical settings to restrict rules extraction to a generic basis of association rules. This one is considered as a reduced set to which we can apply appropriate inference mechanisms to derive redundant rules. In this paper, we introduce a new minimal generic basis MGB of non-redundant association rules based on the augmented Iceberg Galois lattice. The proposed approach involves the inference mechanisms used and a set of experiments applied to several real and synthetic databases. Carried out experiments showed important benefits in terms of reduction in the number of generic rules extracted. We present also a new framework for generating and visually exploring the minimal generic basis MGB.

1 Introduction

Discovering association rules in large datasets is an interesting problem of research in data mining. Its principal objective is to find out correlations between frequents itemsets. An association rule is a strong one when its support (i.e., frequency of the represented pattern) and confidence (i.e., strength of the dependency between the premise and the conclusion) are higher than minimum thresholds fixed by the user (*minsup* and *minconf*). However the number of association rules generated is often huge because many of them are redundant. So, a possible solution to this problem, is to restrict rules extraction to a reduced set containing only non redundant ones which are strictly related to the user's need [1].

This reduced set is known as *generic basis* for association rules, to which we can apply appropriate inference mechanisms to derive all strong redundant association rules. A generic basis can be evaluated by three properties such as *informativity*, *compacity* and *inference mechanism* which will be detailed later.

This paper is organized as follows. Section 2 presents the mathematical background of Formal Concept Analysis (FCA) [2] and describes the association rules mining task. Section 3 deals with the problem of eliminating redundant rules by surveying some of the proposed approaches for constructing generic bases of association rules. In section 4, we present a new minimal generic basis of association rules denoted as MGB. Section 5 consolidates the proposed approach by a set of experiments applied to several real synthetic databases. Finally, we will present VIE_MGB as a new framework for visually generating and exploring MGB and the Iceberg Galois lattice. We were inspired by MIRAGE [3] which allows interactive graphical visualization and exploration of minimal association rules.

2 Mathematical background

In this section, we introduce the mathematical notions based on the FCA.

2.1 Basic notions

Formal context: A formal context is a triplet $\mathcal{K} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$, where $\mathcal{O}$ represents a finite set of objects, $\mathcal{A}$ a finite set of attributes, and $\mathcal{R}$ a binary relation(i.e., $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{A}$). Each couple $(o, a) \in \mathcal{R}$ expresses that the transaction $o \in \mathcal{O}$ contains the attribute $a \in \mathcal{A}$.

We define two functions, summarizing links between subsets of objects and subsets of attributes induced by R, that map sets of objects to sets of attributes and vice versa. Thus, for a set $O \subseteq \mathcal{O}$, we define
$\phi(O) = \{a \mid \forall o,\, o \in O \Longrightarrow (o, a) \in \mathcal{R}\}$; and for $A \subseteq \mathcal{A}$, $\psi(A) = \{o \mid \forall a,\, a \in A \Longrightarrow (o, a) \in \mathcal{R}\}$. Both functions ϕ and ψ form a Galois connection between the sets $\mathcal{P}(\mathcal{A})$ and $\mathcal{P}(\mathcal{O})$ [4]. Consequently, both compound operators of ϕ and ψ are closure operators, in particular, $\omega = \phi \circ \psi$ is a closure operator.

Frequent closed itemset: An itemset $A \subseteq \mathcal{A}$ is said to be closed if $A = \omega(A)$, and is said to be frequent with respect to *minsup* threshold if $support(A) = \frac{|\psi(A)|}{|O|} \geq minsup$.

Formal concept: A formal concept is a couple $c = (O; A)$, where O is called *extent*, and A is a closed itemset, called *intent*. Furthermore, both O and A are related through the Galois connection, i.e., $\phi(O) = A$ and $\psi(A) = O$.

Minimal generator: An itemset $g \subseteq A$ is called minimal generator of a closed itemset A, if and only if $\omega(g) = A$ and $\nexists\, g\prime \subseteq g$ such that $\omega(g\prime) = A$ [5].

Galois lattice: Given a formal context $\mathcal{K}$, the set of formal concepts $\mathcal{C}_\mathcal{K}$ is a complete lattice $\mathcal{L}_c = (\mathcal{C}, \leq)$, called the *Galois (concept) lattice*, when $\mathcal{C}_\mathcal{K}$ is considered with inclusion between itemsets [2] [4]. A partial order on formal concepts is defined as follows $\forall c_1, c_2 \in \mathcal{C}_\mathcal{K}$, $c_1 \leq c_2$ iff $intent(c_2) \subseteq intent(c_1)$, or equivalently $extent(c_1) \subseteq extent(c_2)$. The partial order is used to generate the lattice graph, called *Hasse diagram*, in the following manner: there is an arc (c_1, c_2), if $c_1 \preceq c_2$ where $\preceq$ is the transitive reduction of $\leq$, i.e., $\forall c_3 \in \mathcal{C}_\mathcal{K}$, $c_1 \leq c_3 \leq c_2$ implies either $c_1 = c_3$ or $c_2 = c_3$.

Property 1 *Lattice operator* join *provides the least upper bound (**LUB**) in the concept lattice, and it is defined as follows:*

Let $S \subseteq \mathcal{L}_c$

$$Join(S) = \omega\left(\bigcup_{c \in S} c\right)$$

Iceberg Galois lattice: When only frequent closed itemsets are considered with set inclusion, the resulting structure $(\mathcal{L}_c, \subseteq)$ only preserves the LUB_S, i.e., the join operator. This is called a *join semi-lattice* or *upper semi-lattice*. In the remaining of the paper, such structure is referred to as "*Iceberg Galois Lattice*".

Example 1 *Let us consider the extraction context given by table 1. The associated Iceberg Galois lattice, for minsup* $= \frac{1}{2}$*, is depicted by figure 1. Each node in the Iceberg is represented as couple (closed itemset,support) and is decorated with its associated minimal generators list.*

2.2 Association rules

An association rule represents an implication between frequent itemsets. It is an expression of the form: $R : X \Longrightarrow Y$ [6], in which X and XY are frequent itemsets, and $X \subset XY$. Itemsets X and Y are called, respectively, *premise* and *conclusion* of the rule R. The support of R denoted $supp(R)$, is equal to the support of XY, and its confidence denoted by $conf(R)$ is determined as follows:

R	A	C	D	T	W
1	×	×		×	×
2		×	×		×
3	×	×		×	×
4	×	×	×		×
5	×	×	×	×	×
6		×	×	×	

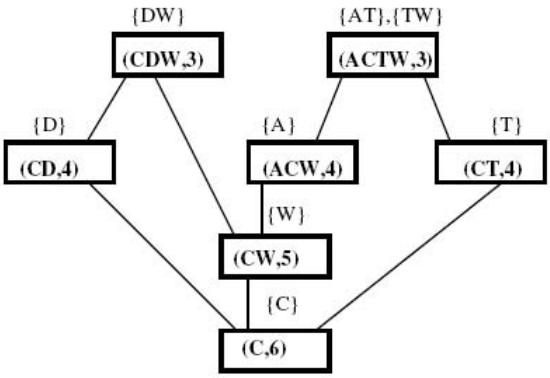

Fig. 1. Iceberg Galois lattice for $minsup = \frac{1}{2}$

$conf(R) = \frac{supp(XY)}{supp(X)}$. If $conf(R) = 1$, then R is an *exact association rule* (ER), otherwise it is called *approximate association rule* (AR).

The process of association rules extraction can be split into two steps [6]:

1. Finding out all frequent itemsets from a given context extraction k for a specified $minsup$ value.
2. Generating all association rules from the frequent itemsets obtained and restricting extraction to rules which confidence satisfies the $minconf$ value.

3 Extraction of non redundant association rules

The problem of the relevance and usefulness of extracted association rules is of primary importance. Indeed, in most real life databases, the set of association rules can rapidly grow to be unwieldy, especially as we lower the $minsup$ value, and among which many are redundant. So, a possible solution to this problem is to restrict extraction of rules to a generic basis of non redundant association rules. Once this generic basis is obtained, all the remaining (redundant) rules can be derived using an appropriate inference mechanism. In an ideal case a generic basis should satisfy the following conditions [7]:

- ***Informativity:*** to exactly determine the support and confidence of redundant association rules.
- ***Deriving redundant rules:*** the derivation axioms should be lossless (should enable derivation of all strong rules) and sound (should forbid derivation of rules that are not strong).
- ***Compacity:*** to have the most reduced set of generic rules allowing the derivation of all strong remaining ones(redundant rules).

In this section, we will overview some representations of strong association rules.

3.1 Minimal non redundant association rules (MNR)

In [5], Bastide et *al.* present a couple of generic bases, one for the exact rules denoted *GB*, and the second for approximate ones denoted *IB*.

Bastide et *al.* consider the following rule-redundancy definition:

Definition 1 *An association rule $R : X \Longrightarrow Y$ is a minimal non redundant association rule iff $\nexists$ an association rule $R_1 : X_1 \Longrightarrow Y_1$ fulfilling the following constraints:*

1. $supp(R) = supp(R_1)$ *and* $conf(R) = conf(R_1)$
2. $X_1 \subset X$ *and* $Y \subset Y_1$

Exact association rules, of the form $R : X \Longrightarrow Y$, are implications between two itemsets X and XY whose closures are identical, i.e., $\omega(X) = \omega(XY)$. Indeed, from the aforementioned equality, we deduce that X and XY belong to the same class of the equivalence relation induced by the closure operator ω on $P(A)$ and then $supp(X) = supp(XY)$ (i.e., $conf(R) = 1$). Bastide et *al.* characterized what they called "the generic basis for exact association rules" (adapting the global implications base of Duquenne and Guigues [8]). The generic basis for exact association rules, that has been proved to be lossless information, is defined as follows:

Definition 2 *Let FC be the set of frequent closed itemsets extracted from a context and, for each frequent closed itemset c, let us denote G_c the set of minimal generators of c. The generic basis for exact association rules is as follows:*

Association rules	*Support*	*Confidence*
$D \Longrightarrow C$	$\frac{2}{3}$	1
$T \Longrightarrow C$	$\frac{2}{3}$	1
$W \Longrightarrow C$	$\frac{5}{6}$	1
$A \Longrightarrow CW$	$\frac{2}{3}$	1
$DW \Longrightarrow C$	$\frac{1}{2}$	1
$TW \Longrightarrow AC$	$\frac{1}{2}$	1
$AT \Longrightarrow CW$	$\frac{1}{2}$	1

Association rules	*Support*	*Confidence*
$T \Longrightarrow ACW$	$\frac{1}{2}$	$\frac{3}{4}$
$A \Longrightarrow CTW$	$\frac{1}{2}$	$\frac{3}{4}$
$D \Longrightarrow CW$	$\frac{1}{2}$	$\frac{3}{4}$
$C \Longrightarrow W$	$\frac{5}{6}$	$\frac{5}{6}$
$C \Longrightarrow T$	$\frac{2}{3}$	$\frac{2}{3}$
$C \Longrightarrow D$	$\frac{2}{3}$	$\frac{2}{3}$

Table 2. Non redundant exact and approximate association rules extracted from the context extraction k for $minsup = \frac{1}{2}$ and $minconf = \frac{2}{3}$

$$GB = \{R : g \Longrightarrow c - g \mid c \in FC \ \wedge g \in G_c \wedge g \neq c\}.$$

The authors also characterized the informative basis for approximate association rules based on extending the family of bases of partial implications of Luxemburger [8], e.g., cover, structural and arborescent bases. The informative basis is defined as follows:

Definition 3 *The informative basis for approximate association rules is given by:*

$$IB = \{R : g \Longrightarrow c - g, c \in FC \wedge g \in G \wedge \omega(g) \subset c\}.$$

With respect to Definitions 2 and 3, we consider that given an Iceberg Galois lattice, representing precedence-relation-based ordered closed itemsets, generic bases of association rules can be derived in a straightforward manner. We assume that in such structure, each closed itemset is "decorated" with its associated list of minimal generators. Hence, AR represent "inter-node" implications, assorted with a statistical information, i.e., the confidence, from a sub-closed-itemset to a super-closed-itemset while starting from a given node in an ordered structure. Inversely, ER are "intra-node" implications extracted from each node in the ordered structure.

Example 2 *Generic bases of exact and approximate association rules that can be drawn from the context k, are respectively depicted in table 2.*

As shown in [7] the couple (GB,IB) form a sound and informative generic basis. However, it produces a very important number of association rules, so this generic basis is not compact.

Association rules	*Support*	*Confidence*
$A \Longrightarrow ACTW$	$\frac{1}{2}$	$\frac{3}{4}$
$T \Longrightarrow ACTW$	$\frac{1}{2}$	$\frac{3}{4}$
$D \Longrightarrow CDW$	$\frac{1}{2}$	$\frac{3}{4}$
$\emptyset \Longrightarrow CD$	$\frac{2}{3}$	$\frac{2}{3}$
$\emptyset \Longrightarrow CT$	$\frac{2}{3}$	$\frac{2}{3}$
$\emptyset \Longrightarrow ACW$	$\frac{2}{3}$	$\frac{2}{3}$

Table 3. Representative basis RB_{Phan} extracted from context extraction k for $minsup = \frac{1}{2}$ and $minconf = \frac{2}{3}$.

3.2 Representative basis ($\mathbf{RB}_{\boldsymbol{Phan}}$)

In [9], the author proposes a new generic basis for association rules denoted RB_{Phan}, and presents its associated inference mechanism. This approach takes as input all frequent itemsets obtained from a given context extraction k by applying an appropriate algorithm such as Apriori [10]. Phan defines the rule-redundancy as follows:

Definition 4 *Let E be the set of all strong association rules extracted from a context extraction k. The rule $r : l_1 \Longrightarrow l_2 \in E$ is a redundant one if and only if there is a rule $r' : l_1' \Longrightarrow l_2' \in E$, and the two following conditions are satisfied:*

1. $l_1' \subseteq l_1$
2. $l_2 \subseteq l_2'$

So the representative basis RB_{Phan} is defined as follows:

Definition 5 *Let $minsup$ and $minconf$ be the support and confidence thresholds for interesting association rules.*

$$RB_{Phan} = \left\{ r : I \Longrightarrow J \mid I \subset J \wedge \nexists r' : I' \Longrightarrow J' \mid I' \subseteq I \wedge J \subseteq J' \right\}$$

Remark 1. RB_{Phan} consists in a redefinition of representative rules denoted as RR and defined in [7]. However, the difference between them is that for RB_{Phan} the premise and the conclusion of a given rule are not necessarily disjoint.

Example 3 *Given the context extraction k depicted by table 1, we present the representative basis extracted for $minsup = \frac{1}{2}$ and $minconf = \frac{2}{3}$ in table 3.*

Once the representative basis is generated, we can derive all strong redundant association rules by applying the inference mechanism defined in [9] as follows:

Definition 6 *Let I, J, K be frequent itemsets and having* RB_{Phan}*, we consider two derivation axioms:*

1. ***Left augmentation:*** *if $I \Longrightarrow JK$ is interesting, then $IJ \Longrightarrow JK$ is interesting.*
2. ***Decomposition:*** *if $I \Longrightarrow I_1 I_2$ is interesting, then $I \Longrightarrow I_1$ and $I \Longrightarrow I_2$ are interesting.*

Moreover, Phan showed that the representative basis is the most reduced set of association rules, from which all remaining rules (redundant) can be derived, so the property of compacity is satisfied. RB_{Phan} is sound and lossless, therefore, this approach presents some drawbacks such as:

1. The representative basis is not informative. The support and confidence of derived rules can not be found.
2. For each rule $R : X \Longrightarrow Y$, we must verify if there is no rule $R^{'} : X^{'} \Longrightarrow Y^{'}$ such as $X^{'} \subseteq X \wedge Y \subseteq Y^{'}$.

3.3 Informative Generic Basis (IGB)

In [11], the authors propose a new informative generic basis for association rules denoted IGB, which takes as input all the frequent closed itemsets. IGB is defined in [11] as follows:

Definition 7 *Let FC be the set of frequent closed itemsets and G_c, the list of minimal generators for a given frequent closed itemset c.*

$$IGB{=}\{R : g_s \Longrightarrow I - g_s \mid I \in FC \wedge g_s \in G_c, c \in FC \wedge c \subseteq I$$
$$\wedge\ conf(R){\geq}minconf \wedge \nexists g^{'} \mid g^{'} \subset g_s \wedge conf(g^{'} \Longrightarrow A - g^{'}) \geq minconf\}$$

Example 4 *Given the context extraction k of table 1, we present the basis* IGB *extracted for $minsup = \frac{1}{2}$ and $minconf = \frac{2}{3}$ in table 4.*

In [11], the authors prove that IGB is informative. In fact, IGB contains all frequent closed itemsets, and the support of a frequent itemset is equal to the support of the smallest frequent closed itemset which contains it. So, the support and the confidence of redundant rules can easily be determined.

In other way, IGB is sound and lossless, and the inference mechanism is composed by three derivation axioms which are [11]:

Association rules	*Support*	*Confidence*
$A \Longrightarrow CTW$	$\frac{1}{2}$	$\frac{3}{4}$
$T \Longrightarrow ACW$	$\frac{1}{2}$	$\frac{3}{4}$
$D \Longrightarrow CW$	$\frac{1}{2}$	$\frac{3}{4}$
$\emptyset \Longrightarrow CD$	$\frac{2}{3}$	$\frac{2}{3}$
$\emptyset \Longrightarrow CT$	$\frac{2}{3}$	$\frac{2}{3}$
$\emptyset \Longrightarrow ACW$	$\frac{2}{3}$	$\frac{2}{3}$
$\emptyset \Longrightarrow CW$	$\frac{5}{6}$	$\frac{5}{6}$
$\emptyset \Longrightarrow C$	1	1

Table 4. IGB extracted from the context extraction k for $minsup = \frac{1}{2}$ and $minconf = \frac{2}{3}$.

1. **Conditional Reflexivity:** if $X \Longrightarrow Y \in$ IGB$\wedge$ $X \neq \emptyset$, then $X \Longrightarrow Y$ is a strong association rule.
2. **Augmentation:** if $X \Longrightarrow Y \in$ IGB, then $X \cup Z \Longrightarrow Y - \{Z\}$ is a strong association rule, $Z \subset Y$.
3. **Conditional Decomposition:** if $R : X \Longrightarrow Y \in$ IGB$\wedge$ $X \neq \emptyset$, then $X \Longrightarrow Z$ is a strong association rule, $Z \subset Y$.

Certainly, IGB is informative, therefore this property is satisfied by the presence of all frequent closed itemsets in the generic basis. So, one can imagine the number of frequent closed itemsets extracted from real databases and consequently the number of generic rules that can be extracted.

4 MGB: A new Minimal Generic Basis

In the previous section, we have studied the properties satisfied by each approach presented, and so, we deduce that, the problem of the construction of generic bases is to find a compromise among the compacity and the informativity of such basis. Then, we propose a new minimal generic basis of association rules that guarantees a *partial informativity* and which means that we can exactly determine the support and confidence of some derived rules, and for others, this approach allows to delimit this measures in an interval.

MGB is defined as follows:

Definition 8 *Given*

1. $\mathcal{L}_c$*: Iceberg Galois lattice decorated by minimal generators and their supports.*

2. c_i*: frequent closed itemset.*
3. $\mathcal{S}$*: set of immediate successors of a given concept* c*.*
4. $\mathcal{G}_{c_i}$*: list of minimal generators of a concept* c_i*.*

$$\text{MGB}=\left\{\begin{matrix} R : g \Longrightarrow c_i - g \mid g \in G_{c_i} \wedge c_i \in \mathcal{L}_c \wedge conf(R) \geq minconf \\ \wedge \nexists\ s \in \mathcal{S} \mid \frac{support(s)}{support(g)} \geq minconf \end{matrix}\right\}$$

4.1 Algorithm for discovery MGB from frequent closed itemsets

In this section, we present the algorithm that extracts the MGB basis directly from Iceberg Galois lattice decorated with its associated list of minimal generators and their supports. It is assumed that for every frequent closed itemset we determine the set of its immediate successors and the list of generators satisfying a $minconf$ constraint.

Algorithm 1 MGB

Input: $\mathcal{L}_c$*: Iceberg Galois lattice decorated by minimal generators and their supports.*

Output: MGB

begin

for each $(c_i \in FC \mid |c_i| > 1)$ *do*

$\mathcal{G}_{c_i} = \{minimal\ generators\ of\ concept\ c^{'} \mid c^{'} \subseteq c \wedge \frac{support(c_i)}{support(c^{'})} \geq minconf\}$

for each generator $g \in \mathcal{G}_{c_i}$ *do*

if $(\nexists s \in \mathcal{S}_{c_i} \mid \frac{support(s)}{support(g)} \geq minconf)$ *then*

MGB $=$MGB $\cup R : g \rightarrow c_i - g$

$\mathcal{G}_{c_i} = \mathcal{G}_{c_i} - \{g^{'} \mid g \subseteq g^{'}\}$

else

$\mathcal{G}_{c_i} = \mathcal{G}_{c_i} - \{g\}$

end

Example 5 *Given the context extraction shown by table 1, we will illustrate in table 5 the* MGB *generic basis for* $minsup = \frac{1}{2}$ *and* $minconf = \frac{2}{3}$*.*

As we have shown, MGB requires to determine the list of all immediate successors of a given concept c, thus, we will present how to construct this list. So, we will explore the Iceberg Galois lattice considering the set inclusion and the partial order defined between the different frequent closed itemsets. In fact, this process is composed of two steps, which are:

1. Generating the list of all successors of a given concept c.
2. Using this list, we determine the immediate successors of c.

Association rules	*Support*	*Confidence*
$A \Longrightarrow CTW$	$\frac{1}{2}$	$\frac{3}{4}$
$T \Longrightarrow ACW$	$\frac{1}{2}$	$\frac{3}{4}$
$W \Longrightarrow AC$	$\frac{2}{3}$	$\frac{4}{5}$
$C \Longrightarrow AW$	$\frac{2}{3}$	$\frac{2}{3}$
$C \Longrightarrow T$	$\frac{2}{3}$	$\frac{2}{3}$
$C \Longrightarrow D$	$\frac{2}{3}$	$\frac{2}{3}$
$D \Longrightarrow CW$	$\frac{1}{2}$	$\frac{3}{4}$

Table 5. The generic basis MGB obtained from the context extraction k for $minsup = \frac{1}{2}$ and $minconf = \frac{2}{3}$

4.2 Deriving redundant association rules

For deriving all strong redundant association rules, we propose to adopt two derivation axioms of the inference mechanism defined in [11].

1. **Augmentation:** if $X \Longrightarrow Y \in$ MGB, then $X \cup Z \Longrightarrow Y - \{Z\}$ is a strong association rule, $Z \subset Y$.
2. **Conditional Decomposition:** if $R : X \Longrightarrow Y \in$ MGB, then $X \Longrightarrow Z$ is a strong association rule, $Z \subset Y$.

This inference mechanism has been proved in [11] to be sound and complete.

Remark 2. It is important to mention that for deriving all approximate association rules, the order of applying the inference axioms is important. Indeed, we have to apply first the Augmentation axiom and next apply the Conditional Decomposition axiom on the resulting set of association rules, i.e., set of approximate generic rules and those derived by the Augmentation axiom.

In fact, MGB is not informative. The support and confidence of derived rules cannot be found. For example we cannot find the confidence of the rule $AT \Longrightarrow CW$ derived from $A \Longrightarrow CTW$ because the support of AT is unknown.

Nevertheless, the *non-informativity* problem can be partially resolved by introducing a new concept denoted "*Partial informativity*". In fact, it means that we can exactly determine the support and confidence of some derived rules, and for others, this approach allows to delimit this measures in an interval.

Proposition 1. *Let* $R : X \Longrightarrow Y \in$MGB*,* A *and* B *two integers, if we apply the inference mechanism, we obtain rules of the following forms:*

1. Augmentation: $R_1 : XZ \Longrightarrow Y - Z$

A=B	A≠B
Augmentation	**Augmentation**
R1.supp=R.supp	R1.supp=R.supp
R1.conf=R.supp/supp(XZ)	R1.conf∈[R.conf,1]
Conditional Decomposition	**Conditional Decomposition**
R2.supp=supp(XZ)	B $\leq$ R2 .supp $\leq$ A
R2.conf=supp(XZ)/supp(X)	R2.conf∈ [R.conf,1]

Table 6. Support an confidence of derived rules

2. *Conditional Decomposition:* $R_2 : X \Longrightarrow Z$
   (a) $support(XZ) \leq min\,\{support(i_1), support(i_2), ..., support(i_k) \mid i_i \subset XZ\}$, *A, where* i_i *is a frequent subset of* XZ.
   (b) $support(XZ) \geq max\{support(I_1), support(I_2), ..., support(I_k)\}$, *B, where* I_i *is a frequent closed itemset containing* XZ *and existing in* MGB.

Remark 3. if $A = B$, then we have $supp(XZ) = A$.

So, table 6 illustrates how we determine support and confidence of these rules.

Proof. a: Since the support of a frequent itemset is less than the support of all its subsets, then we deduce that $support(XZ)$ is less than the smallest support of all its subsets.

b: By same, the support of a frequent itemset is higher than the support of all its supersets, so $support(XZ)$ is higher than the support of the smallest frequent closed itemset which contains it and existing in MGB.

c: If we apply the augmentation axiom, the generic rule's base and the redundant rule's base are similar, so $support(R_1) = support(R)$.

d: If we apply the conditional decomposition axiom, then the derived rule's base is XZ, and so $support(R_2) = support(XZ)$

Remark 4. Before deriving the redundant rules, we determine the support of all frequent itemsets which are the antecedent of generic rules.

Example 6 *Given the generic rule* $R : C \Longrightarrow AW$, $supp(R) = \frac{2}{3}$ *and* $conf(R) = \frac{2}{3}$, *so,* $supp(C) = \frac{supp(R)}{conf(R)} = 1$.

In table 7, we summarize the properties of the reviewed association rules representations. These ones contain only rules whose bases are frequent closed itemsets and whose antecedent are frequent generators. Let us consider the following notations:

Generic basis	Intended derivable rules	Infer. mechanism	Lossless	Sound	Informative
RB_{Phan}	AR	$\mathcal{LA}+\mathcal{D}$	yes	yes	no
IGB	AR	$\mathcal{CR}+\mathcal{A}+\mathcal{CD}$	yes	yes	yes
(GB,IB)	certainAR+approxAR	$\mathcal{A}+\mathcal{CD}$	yes	yes	yes
scMGB	AR	$\mathcal{A}+\mathcal{CD}$	yes	yes	no

Table 7. Summary of Association Rules Representations

Database	Nature	#transactions	#items	$\|freq-clos\|$	$\|max-clos\|$
Retail	sparse	88162	16470	580	284
Mushrooms	dense	8124	120	427	63
Connect	dense	67557	130	810	98
Chess	dense	3196	76	1194	71

Table 8. Dataset characteristics

- $\mathcal{AR}$*:* strong association rules
- $\mathcal{LA}$: left augmentation
- $\mathcal{D}$: decomposition
- $\mathcal{CR}$: conditional reflexivity
- $\mathcal{A}$: augmentation
- $\mathcal{CD}$: conditional decomposition

5 Visual implementation and experimental evaluation

All experiments described below were performed on a 3,06GHz PC with 512MB of memory, running Windows XP. The algorithm MGB was coded in JAVA.

Table 8 shows characteristics of real and synthetic datasets used in our evaluation.

Table 9 shows the experimental results. The last column consists of the number of all strong association rules extracted.

5.1 Experimental results

We compare MGB'results with those concerning RB_{Phan}, IGB and (GB,IB) which are detailed in [11].

Database	Minsup	Minconf	MGB	RB_{Phan}	IGB	(GB,IB)	AR
Retail	0.5%	0.5%	435	284	580	1382	1382
		1%	402	459	757	1334	1334
		10%	232	214	427	770	770
		50%	304	305	353	438	438
		100%	0	0	0	0	0
Connect	95%	95%	635	97	809	25336	77816
		96%	852	1003	2140	23780	73869
		97%	1403	1178	2438	18470	60101
		98%	1033	1404	2598	11717	41138
		99%	1386	1667	2284	5250	19967
		100%	682	682	682	682	2260
Mushrooms	30%	30%	332	63	427	7623	94894
		50%	366	318	967	5761	79437
		70%	364	429	966	4520	58010
		90%	498	519	799	2159	24408
		100%	557	558	558	557	8450
Chess	87%	87%	440	71	1194	31538	42740
		89%	519	430	2423	29704	40451
		91%	627	486	2655	26147	36098
		93%	793	616	2766	21350	29866
		95%	671	768	2754	14373	20312
		100%	342	342	342	342	342

Table 9. Number of generic association rules

– For some $minconf$ values, MGB is the most reduced set of generic association rules:
  - In fact, concerning the representative basis RB_{Phan}, for each frequent closed itemset c, if $support(c) \geq minconf$, then we have a factual rule of the form $\emptyset \Longrightarrow c$.
  - However, for MGB and for the same $minconf$ value, if the set of generators of c is empty, the algorithm do not generate any rule.
– If we consider the database *Retail*, one can observe that for $minconf = 1$ the number of association rules is equal to 0, so we can say that all the frequent closed itemsets have only one minimal generator which overlaps the associated closed itemset.
– By setting $minsup = minconf$
  - For IGB, the number of association rules is equal to the number of frequent closed itemsets. In fact, the support of all the frequent closed itemsets is higher or equal than the $minconf$ value.
  - The number of association rules in RB_{Phan} is equal to the number of maximal frequent closed itemsets,
– For Mushrooms dataset we can observe that the bases IGB and RB_{Phan} have one more association rule than MGB and (GB,IB), because of the presence of the item "85" in all transactions which produces a factual rule in IGB and RB_{Phan} of the form $\emptyset \Longrightarrow 85$.

5.2 VIE_MGB: Visual Interactive Exploration of Minimal Generic Basis

In this section, we present VIE_MGB, a new framework for an interactive graphical visualization and exploration of MGB and the Iceberg Galois lattice. VIE_MGB allows to display generic rules and to derive all strong redundant association rules with their supports and confidences. The database of interest has previously been mined, using an efficient algorithm for mining frequent closed itemsets and their minimal generators. Once these closed itemsets are mined, they are taken as input.

In fact, we have two independent processes executed by VIE_MGB, which are:

1. Generating MGB.
2. Constructing the Iceberg Galois lattice and to visualize redundant rules.

The visual interactive framework for exploring and visualizing minimal generic basis is depicted in figure 2.

To generate MGB As we have shown below, this tool takes as input a previously stored file that contains all the frequent closed itemsets, their minimal generators and supports extracted from a given dataset, then the user must specify a *minconf* value, thus VIE_MGB generates a text file which represents MGB containing all the generic association rules with their supports and confidences.

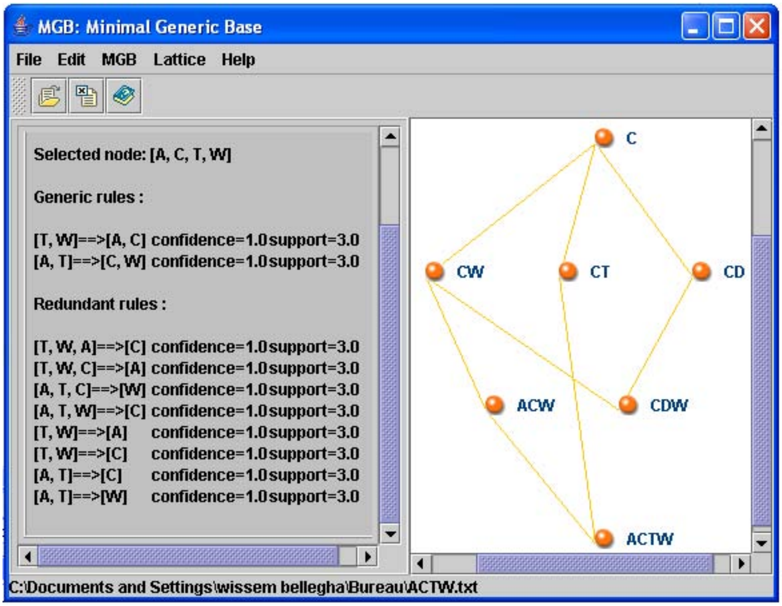

Fig. 2.

Interactive lattice and rule exploration In other way, from the closed itemsets, VIE_MGB creates the Iceberg Galois lattice as shown in figure 2. Each closed itemset is represented by a node, and there is a link connecting two nodes if they are related by a set inclusion relation and there is no intermediate set between them. Once the lattice is displayed on the panel, one can generate the redundant association rules as follows:

The user can click on a chosen node on the lattice, then VIE_MGB displays on the text area the generic rules corresponding to this closed itemset and all strong redundant association rules derived from the generic ones with their supports and confidences.

Remark 5. When generating the redundant association rules, each frequent itemset is mapped to the smallest frequent closed itemset which contains it to determine the support of the derived rule and its confidence. So, the problem of the informativity of the generic basis is completely resolved.

6 Conclusion

In this paper, we have presented theoretical foundations of generic bases of association rules and we have pointed out the properties satisfied by some approaches. We have proposed MGB which is a new minimal generic basis of associations rules and its associated inference mechanism. We have also introduced the concept of "partial informativity".

This new approach is consolidated by a set of experiments using real and synthetic databases showing important benefits in terms of reduction in the number of association rules presented to the user. Finally, we have proposed a new framework denoted VIE_MGB allowing a visual interactive exploration of MGB.

References

1. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Intelligent structuring and reducing of association rules with formal concept analysis. In: Proceedings of KI'2001 conference, Lecture Notes in Artificial Intelligence 2174, Springer-verlag. (2001) 335–350
2. Ganter, B., Wille, R.: Formal Concept Analysis. Springer-Verlag, Heidelberg (1999)
3. Zaki, M., Phoophakdee, B.: MIRAGE: A framework for mining, exploring and visualizing minimal association rules. Rpi cs dept technical report (2003)
4. Barbut, M., Monjardet, B.: Ordre et classification. Algèbre et Combinatoire. Hachette, Tome II (1970)
5. Bastide, Y., Pasquier, N., Taouil, R., Lakhal, L., Stumme, G.: Mining minimal non-redundant association rules using frequent closed itemsets. In: Proceedings of the International Conference DOOD'2000, Lecture Notes in Computer Sciences, Springer-verlag. (2000) 972–986
6. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between sets of items in large Databases. ACM SIGMOD Records (1993) 207–216
7. Kryszkiewicz, M.: Concise representations of association rules. In: Proceedings of Pattern Detection and Discovery, ESF Exploratory Workshop, London, UK. (2002) 92–109

8. Duquenne, J., Guigues, J.: Famille minimale d'implications informatives résultant d'un tableau de données binaires. Mathématiques et Sciences Humaines **95** (1986) 5–18
9. Luong, V.P.: Raisonnement sur les rgles d'association. In: In Proceedings 17me Journe Bases de Donnes Avances BDA'2001, Agadir(Maroc), Cpadus Edition. (2001) 299–310
10. Agrawal, R., Skirant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Databases. (1994) 478–499
11. Gasmi, G., BenYahia, S., Nguifo, E.M., Slimani, Y.: A new informative generic base of association rules. In: In Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-05),Hanoi,Vietnam. (2005) 81–90