A new Informative Generic Base of Association Rules

Gh. Gasmi[1], S. Ben Yahia[1;2], E. Mephu Nguifo[2], and Y. Slimani[1]

[1]Départment des Sciences de l'Informatique, Faculté des Sciences de Tunis
Campus Universitaire, 1060 Tunis, Tunisie.
{sadok.benyahia,yahya.slimani}@fst.rnu.tn
[2]Centre de Recherche en Informatique de Lens-IUT de Lens
Rue de l'Université SP 16, 62307 Lens Cedex
mephu@cril.univ-artois.fr

Abstract. The problem of the relevance and the usefulness of extracted association rules is becoming of primary importance, since an overwhelming number of association rules may be derived from even reasonably sized real-life databases. In this paper, we introduce a novel generic base of association rules, based on the Galois connection semantics. The novel generic base is sound and informative. We also present a sound axiomatic system, allowing to derive all association rules that can be drawn from an extraction context.

1 Introduction

Data mining has been extensively addressed for the last years, particularly the problem of discovering association rules. These latter aim at exhibiting correlations between data items (or attributes), whose interestingness is assessed by statistical metrics. However, an unexploited huge amount of association rules is drawn from real-life databases. This drawback encouraged many research issues, aiming at finding the minimal nucleus of relevant knowledge can be extracted from several thousands of highly redundant rules. Various techniques are used to limit the number of reported rules, starting by basic pruning techniques based on thresholds for both the frequency of the represented pattern (called the *support*) and the strength of the dependency between premise and conclusion (called the *confidence*). More advanced techniques that produce only a limited number of the entire set of rules rely on closures and Galois connections [1–3]. These formal concept analysis (FCA) [4] based techniques have in common a feature, which is to present a better trade-off between the size of the mining result and the conveyed information than the "frequent patterns" algorithms. Finally, works on FCA have yielded a row of results on compact representations of closed set families, also called *bases*, whose impact on association rule reduction is currently under intensive investigation within the community [1, 2, 5].

Once these generic bases are obtained, all the remaining (redundant) rules can be derived "easily". In this context, little attention was paid to reasoning

© V. Snášel, R. Bělohlávek (Eds.): CLA 2004, pp. 67–79, ISBN 80-248-0597-9.
VŠB – Technical University of Ostrava, Dept. of Computer Science, 2004.

from generic bases comparatively to the battery of papers to define them. Essentially, they were interested in defining syntactic mechanisms for deriving rules from generic bases.

In this paper, we introduce a novel generic base of association rule, which is sound and informative. The soundness property assesses the "syntactic" derivation, since it ensures that all association rules can be derived from the generic base. The informativeness property ensures that the support and confidence of a derivable rule can be exactly determined.

The remainder of the paper is organized as follows. Section 2 introduces the mathematical background of FCA and its connection with the derivation of (non-redundant) association rule bases. Section 3 presents the related work on defining and reasoning from generic bases of association rules. In section 4, we introduce a novel, sound and informative generic base of association rules. We also provide a set of inference axioms, for deriving association rules and we we prove its soundness. Section 5 concludes this paper and points out future research directions.

2 Mathematical background

In the following, we recall some key results from the Galois lattice-based paradigm in FCA and its applications to association rules mining.

2.1 Basic notions

In the rest of the paper, we shall use the theoretical framework presented in [4]. In this paragraph, we recall some basic constructions from this framework.

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 (or transactions), $\mathcal{A}$ is a finite set of attributes and $\mathcal{R}$ is a binary (incidence) 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 $a \in \mathcal{A}$. Within a context (c.f., Figure 1 on the left), objects are denoted by numbers and attributes by letters.

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 $\phi(O) = \{a \mid \forall o, o \in O \Rightarrow (o, a) \in \mathcal{R}\}$; and for $A \subseteq \mathcal{A}$, $\psi(A) = \{o \mid \forall a, a \in A \Rightarrow (o, a) \in \mathcal{R}\}$. Both functions ϕ and ψ form a Galois connection between the sets $\mathcal{P}(\mathcal{A})$ and $\mathcal{P}(\mathcal{O})$ [6]. Consequently, both compound operators of ϕ and ψ are closure operators, in particular $\omega = \phi \circ \psi$ is a closure operator.

In what follows, we introduce the frequent closed itemset [3], since we may only look for itemsets that occur in a sufficient number of transactions.

[3] Itemset stands for a set of items

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 supp(A)= $\frac{|\psi(A)|}{|\mathcal{O}|} \geq minsup$.

Formal Concept: A formal concept is a pair $c = (O, A)$, 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 $g \subseteq \mathcal{A}$ is *called minimal generator* of a closed itemset A, if and only if $\omega(g) = A$ and $\nexists g' \subseteq g$ such that $\omega(g') = A$ [1]. The closure operator ω induces an equivalence relation on items power set, i.e., the power set of items is partionned into disjoint subsets (also called *classes*). In each distinct class, all elements are equal support value. The minimal generator is the smallest element in this subset, while the closed itemset is the largest one. Figure 1(Right) sketches sample classes of the induced equivalence relation from the context $\mathcal{K}$.

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 [4, 6]. A partial order on formal concepts is defined as follows $\forall\ c_1, c_2 \in \mathcal{C}_\mathcal{K}$, $c_1 \leq c_2$ iif $intent(c_2) \subseteq intent(c_1)$, 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 $c_2 = c_3$.

Iceberg Galois lattice: When only frequent closed itemsets are considered with set inclusion, the resulting structure $(\hat{\mathcal{L}}, \subseteq)$ only preserves the LUBs, i.e., the joint 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 Figure 1 (Left). The associated Iceberg Galois lattice, for minsup=2, is depicted by Figure 1(Bottom)[4]. Each node in the Iceberg is represented as couple (closed itemset; support) and is decorated with its associated minimal generators list.

In the following, we present the general framework for the derivation of association rules, then we establish its important connexion with the FCA framework.

2.2 Derivation of association rules

Let $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$ be a set of m distinct items. A transaction T, with an identifier further called *TID*, contains a set of items in $\mathcal{I}$. A subset X of $\mathcal{I}$ where $k = |X|$ is referred to as a $k-$itemset (or simply an itemset), and k is called the length of X. A transaction database, say $\mathcal{D}$, is a set of transactions, which can be easily transformed in an extraction context $\mathcal{K}$. The number of transactions of $\mathcal{D}$ containing the itemset X is called the support of X, i.e.,

[4] We use a separator-free form for sets, e.g., AB stands for $\{A, B\}$.

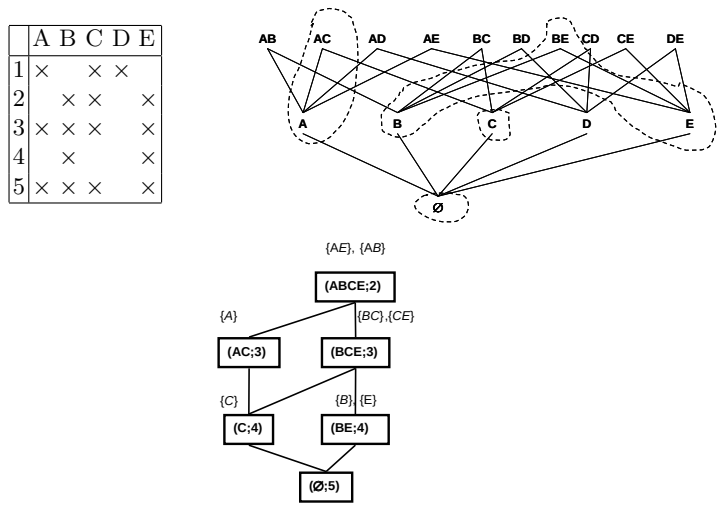

Fig. 1. Left: Extraction context $\mathcal{K}$ **Right:** Associated sample of equivalence relation classes **Bottom:** Associated Iceberg Galois lattice for minsup =2.

supp(X)=| $\{T \in \mathcal{D} \mid X \subseteq T\}$ |. An itemset X is said to be frequent when support(X) reaches at least the user-specified minimum threshold minsup.

Association rule derivation is achieved from a set F of frequent itemsets in an extraction context $\mathcal{K}$, for a minimal support *minsup*. An association rule R is a relation between itemsets of the form $R : X \Rightarrow (Y - X)$, in which X and Y are frequent itemsets, and $X \subset Y$. Itemsets X and $(Y - X)$ are called, respectively, *premise* and *conclusion* of the rule R. The valid association rules are those whose the strength metric conf(R)= $\frac{|XY|}{|X|}$ is greater than or equal to the minimal threshold of confidence minconf. If conf(R)=1 then R is called *exact association rule (ER)*, otherwise it is called *approximative association rule (AR)*.

3 Related work on generic bases of association rules

The problem of the relevance and usefulness of extracted association rules is of primary importance. Indeed, in most real life databases, thousands and even millions of high-confidence rules are generated among which many are redundant. This problem encouraged the development of tools for rule classification according to their properties, for rule selection according to user-defined criteria, and for rule visualization. In this paper, we are specially interested in the lossless information reduction, which is based on the extraction of a generic subset of all association rules, called *base*, from which the remaining (redundant) association rules may be derived. The concept of rule redundancy can be considered as follows [7]:

Definition 1. *Let $\mathcal{AR}_{\mathcal{K}}$ be the set of association rules that can be drawn from a context $\mathcal{K}$. A rule R: $X \stackrel{c}{\Rightarrow} Y$* [5] *$\in \mathcal{AR}_{\mathcal{K}}$ is considered as redundant (or derivable) with respect to (from) R_1: $X_1 \stackrel{c}{\Rightarrow} Y_1$ if R fulfils the following conditions:*

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

Exact association rules, of the form R : $X \Rightarrow (Y)$, 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 $\mathcal{P}(\mathcal{A})$ and then supp(X)= supp(XY) (i.e., conf(R) = 1).

3.1 Work of Bastide *et al.*

Bastide *et al.* characterized what they called "*the generic basis for exact association rules*" (adapting the global implications base of *Duquenne and Guigues* [8]).

Definition 2. *Let FC be the set of frequent closed itemsets extracted from the context and, for each frequent closed itemset c, let us denote $\mathcal{G}_c$ the set of minimal generators of c. The generic basis for exact association rules, called $\mathcal{GBE}$, is defined as follows:*
$\mathcal{GBE}$ $=\{R : g \Rightarrow (c - g) \mid c \in \mathcal{FCI}$ *and* $g \in \mathcal{G}_c$ *and* $g \neq c\}$.

The authors also characterized a generic base of approximate association rules, based on extending the partial implications base of *Luxenburger* [9]. The $\mathcal{GBA}$ base is defined as follows :

Definition 3. *The generic base of approximate association rules given by :* $\mathcal{GBA}$ $= \{R|\ R : X \Rightarrow Y,\ Y \in \mathcal{FCI}$ *and* $\omega(X) \leq Y$ *and c=conf(R)* $\geq$ *minconf and supp(Y)* $\geq$ *minsup*$\}$.

As pointed out in [10], the couple ($\mathcal{GBE}$,$\mathcal{GBA}$) of generic bases form a sound and informative generic base. 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. For example, Generic bases of exact and approximative association rules that can be drawn from the context $\mathcal{K}$, are respectively depicted in Figure 2 (Left and Right).

In [7], the authors presented sound inference axioms for both ($\mathcal{GBE}$ and $\mathcal{GBA}$) bases, permitting to derive all association rules from generic bases of association rules.

[5] When conf(R: $X \stackrel{c}{\Rightarrow} Y$) =1, then c is omitted, i.e., R is written as R: $X \Rightarrow Y$.

Rule #	Implication
R_1	E⇒B
R_2	B⇒E
R_3	A⇒C
R_4	BC⇒E
R_5	CE⇒B
R_6	AB⇒CE
R_7	AE⇒BC

Rule #	Implication	Rule #	Implication
R_8	$C \overset{.75}{\Rightarrow} A$	R_{13}	$B \overset{.75}{\Rightarrow} CE$
R_9	$C \overset{.75}{\Rightarrow} BE$	R_{14}	$E \overset{.5}{\Rightarrow} ABC$
R_{10}	$C \overset{.5}{\Rightarrow} ABE$	R_{15}	$B \overset{.5}{\Rightarrow} ACE$
R_{11}	$A \overset{.66}{\Rightarrow} BCE$	R_{16}	$BC \overset{.66}{\Rightarrow} AE$
R_{12}	$E \overset{.75}{\Rightarrow} BC$	R_{17}	$CE \overset{.66}{\Rightarrow} AB$

Fig. 2. Generic bases drawn from the context $\mathcal{K}$ **Left:** the $\mathcal{GBE}$ base **Right:** the $\mathcal{GBA}$ base.

3.2 Work of Phan

In [11], the author presented a definition of a generic base and the associated derivation axioms. The author broke the "monopoly" of the FCA based-semantics related work for deriving generic bases. Indeed, the presented approach to derive the generic base takes as input the set of frequent itemsets (and not closed) yield for example by the Apriori algorithm. Another peculiarity of the proposed generic base is that it is composed of association rules, in which the respective premise and conclusion parts are not necessarily disjoint.

In what follows, we present the GB_{Phan} algorithm, whose pseudo-code is given by Algorithm 1.1, permitting to derive the GB_{Phan} generic base.

Example 2. Let us consider the extraction context given by Figure 1 (Left). Then, in what follows the running process of the GB_{Phan} algorithm for minsup $=\frac{2}{5}$ and minconf $=\frac{1}{2}$. The set $\mathcal{R}$ is initialized with the set of frequent itemsets in a decreasing order, i.e., $\mathcal{R}$= {ABCE, ABC, ABE, ACE, BCE, AB, AC, AE, BC, BE, CE, A, B, C, E}. For J = ABCE, we have supp(ABCE) = $\frac{2}{5} \leq$ minconf. Then, we have : L_{ABCE} = A, B, C, E, AB, AC, AE, BC, BE, CE, ABC, ABE, ACE, BCE. For I = A, we have $\frac{|ABCE|}{|A|} = \frac{2}{3} \geq$ minconf. In this case, the generic rule A⇒ ABCE is added to the generic base and from L_{ABCE} we delete all element including A, i.e., AB, AC, AE, ABC, ABE, ACE, and A. Therefore, the set L_{ABCE} equals {B, C, E, BC, CE, BCE}. Next for I= B, we have $\frac{|ABCE|}{|B|}=\frac{1}{2}$, we have to generate the generic rule B⇒ ABCE and to delete BC, C and BCE from L_{ABCE}. Therefore, the set L_{ABCE} is equal to {C, E, CE}. For I= C, we have $\frac{supp(ABCE)}{supp(C)} = \frac{1}{2}$, then we generate the generic rule C⇒ ABCE and we delete CE and C from L_{ABCE}. Finally for I= E, we have $\frac{|ABCE|}{|E|}=\frac{1}{2}$. Hence, we generate the generic rule E⇒ABCE and E is removed from L_{ABCE}. Next, we have to remove from the set $\mathcal{R}$, the following itemsets ABC, ABE, ACE, AB and AE. The set $\mathcal{R}$ is found to be equal to {BCE, AC, BC, BE, CE, A, B, C, E}. Then, for J= BCE, we have |BCE|= $\frac{3}{5} \geq$ minconf. In this case, we have to generate the generic rule $\emptyset$ ⇒BCE and we delete BCE, BC, BE, CE, B, C and E from $\mathcal{R}$. Therefore, $\mathcal{R}$= {AC, A} and for J= AC we have |AC|= $\frac{3}{5} \geq$ minconf. Then, we have to generate the generic rule $\emptyset$ ⇒ AC. After the removal of AC and A from

Algorithm 1.1: GB_{Phan}

Data:

- $\mathcal{R}$= {frequent itemsets in a decreasing order}
- minsup
- minconf

Result: GB_{Phan}: generic base

begin

  foreach $J \in \mathcal{R}$ *of size i from m down to 1*| m= |*largest frequent itemset*| **do**

    if *support(J)* $\geq$ *minconf* **then**

      GB_{Phan}=GB_{Phan}∪ {$\emptyset \Rightarrow$J }

      $\mathcal{R}$=$\mathcal{R}$-{$J'|J' \subseteq J$ }

    else

      L_J= {all nonempty subset of J in an ascendant order }

      foreach $I \in L_J$ *of size k from 1 to i-1* **do**

        if $\nexists\ I' \Rightarrow J' \in GB_{Phan} \mid I' \subseteq I \wedge$ *and* $J \subseteq J' \wedge \frac{|J|}{|I|} \geq$ *minconf* **then**

          GB_{Phan}=GB_{Phan}∪ $\{I \Rightarrow J\}$

          L_J=L_J-{ $I'|\ I \subset I' \subset J$}

        L_J=L_J- I

      $\mathcal{R}$= $\mathcal{R}$- {J′ ⊆ J| |J′|=|J|}

end

$\mathcal{R}$, the set $\mathcal{R}$ is found to be empty and the algorithm terminates. The resulting generic base is depicted by Table 1.

Implication	**Support**	**Confidence**
A⇒ABCE	$\frac{2}{5}$	$\frac{2}{3}$
B⇒ABCE	$\frac{2}{5}$	$\frac{1}{2}$
C⇒ABCE	$\frac{2}{5}$	$\frac{1}{2}$
E⇒ABCE	$\frac{2}{5}$	$\frac{1}{2}$
$\emptyset \Rightarrow$BCE	$\frac{4}{5}$	$\frac{3}{5}$
$\emptyset \Rightarrow$AC	$\frac{4}{5}$	$\frac{3}{5}$

Table 1. Generic Base GB_{Phan}.

For the derivation of association rules, the author presented a sound axiomatic system composed of the following two axioms:

A1:Left augmentation If I⇒ JK ∈ GB_{Phan}, then IJ⇒JK is a derivable valid rule.

A2:Decomposition If I ⇒ JK∈ GB_{Phan}, then I⇒J and I⇒ K are derivable valid rules.

For example, from $\emptyset \Rightarrow$AC and using the above mentioned axioms, it is possible to derive A$\Rightarrow$ AC, C$\Rightarrow$ AC and A$\Rightarrow$ C.

However, two drawbacks may be addressed. At first, only frequent itemsets are used to define the generic base, and one can imagine the length and the number of such frequent itemsets that could be derived from correlated extraction contexts. Second, the presented generic base is not informative, i.e., it may exist a derivable rule for which it is impossible to determine exactly both its support and confidence. For example, the association rule BE$\Rightarrow$C is derivable from the generic rule E$\Rightarrow$ABCE. However, it is impossible to derive the exact confidence and support of the derivable rule from the GB_{Phan} generic base.

3.3 Work of Kryszkiewicz

In [10], the author introduced another syntactic derivation operator, called the *cover*, defined as follows:

Cover(X $\Rightarrow$Y) = $\{X \cup Z \Rightarrow V \mid Z, V \subseteq Y \wedge Z \cap V = \emptyset \wedge V \neq \emptyset\}$.

The number of the derived rules from a rule R: X $\Rightarrow$ Y is equal to |Cover(R)|=3^m-2^m, where |Y|=m. For a derived rule R', we have conf(R')$\geq$ *max* { conf(R_i)| R'$\in$ Cover(R_i)} and supp(R')$\geq$ *max* { supp(R_i)| R'$\in$ Cover(R_i)}. In order to determine whether a rule R':X'$\Rightarrow$Y' belongs to the cover of a rule R:X$\Rightarrow$Y, we have to check that $X \subseteq X'$ and $Y' \subset Y$.

The author proposed a minimal base of rules called *representative association rules* ($\mathcal{RR}$) defined as follows: $\mathcal{RR}$={R $\in \mathcal{AR} \mid \nexists$ R'$\in \mathcal{AR}$, R$\neq$R'and R $\in \mathcal{C}$(R')} where $\mathcal{AR}$ is the set of all valid association rules.

As pointed out in [10], using the cover operator as a rule inference mechanism, $\mathcal{RR}$ is lossless, sound but it is not informative. However, the author claimed that the couple ($\mathcal{GBE}$,$\mathcal{GBA}$) forms a lossless, sound and informative generic base, by using the cover operator as inference mechanism.

On the other hand, the author showed that given the following bases:

- The $\mathcal{DG}$ *Duquenne-Guigues* basis for exact rules, i.e., $\mathcal{DG}$={R:X$\Rightarrow\omega(X)$-X |X $\in\mathcal{FP}$}, where $\mathcal{FP}$ the pseudo-closed itemsets.
- the $\mathcal{PB}$ Proper basis for approximative rules, i.e., $\mathcal{PB}$={R:X$\Rightarrow$Y-X |X,Y $\in\mathcal{FCI}$ and $X \subset Y$ and conf(R)$\leq$ minconf}

then the couple($\mathcal{PB}$,$\mathcal{DG}$) forms a lossless, sound and informative base, by using the conjunction of *closure-closure* inference rule [10] and Armstong's axioms [12] as inference mechanism. Clearly, the drawback of the solution proposed is the considerable increase in the size of base in order to be sound and informative.

4 New generic base

Intuitively, we are looking for defining a new informative generic base providing the "maximal boundaries" in which stand all the derivable rules. Indeed, a derivable rule cannot present a smaller premise than that of its associated

generic rule, i.e., from which it can be derived. On the other hand, a derivable rule cannot present a larger conclusion than that of its associated generic rule. In what follows, we present the definition of the introduced $\mathcal{IGB}$ generic base.

Definition 4. *The generic base $\mathcal{IGB}$ of association rules given by : $\mathcal{IGB}$ = {R : $g_s \Rightarrow A\text{-}g_s \mid A \in \mathcal{FCI} \wedge \omega(g_s) \leq A \wedge conf(R) \geq minconf \wedge \nexists\ g'$ such that $g_s \subset g'$ and $conf(g' \Rightarrow A\text{-}g') \geq minconf$ }.*

Proposition 1. *Let I a frequent closet itemset. If $|I| \geq minconf$, then the generic rule that can be drawn from I is $\emptyset \Rightarrow I$.*

Proof. since conf($\emptyset \Rightarrow$I)=supp(I), then the generic rule $\emptyset \Rightarrow$I is valid. It is also the largest rule that can be drawn from the frequent closed itemset I. Indeed, $\nexists$ R_1: $X_1 \Rightarrow Y_1$ such that $X_1 \subset \emptyset$ and I $\subseteq Y_1$.

4.1 The $\mathcal{IGB}$ generic base construction

In what follows, we present the IGB algorithm, whose pseudo-code is given by Algorithm1.2, permitting to construct the introduced generic base of association rules. The IGB algorithm is based on Proposition 1. So, it considers the set of frequent closed itemsets $\mathcal{FCI}$. For each closed itemset I, it checks whether its support is greater than or equal to minconf. If it is the case, then we generate the generic rule $\emptyset \Rightarrow$I. Otherwise, it has to look for the smallest minimal generator, say g_s, associated to a frequent closed itemset subsumed by I, and generates the generic rule $g_s \Rightarrow$I-g_s.

Example 3. Let us consider the extraction context given by Figure 1 (Left). Then, in what follows the running process of the IGB algorithm for minsup $=\frac{2}{5}$ and minconf $= \frac{1}{2}$. For I = ABCE, we have |ABCE|$= \frac{2}{5} < \frac{1}{2}$, L_{ABCE} = {C, A, B, E, BC, CE, AE, AB}. Since $\frac{|ABCE|}{|C|} = \frac{1}{2}$, we have to generate the generic rule C$\Rightarrow$ABE and to remove BC, CE and C from L_{ABCE}. Next, we have $\frac{|ABCE|}{|A|}$ $= \frac{2}{3} > \frac{1}{2}$. Then, we generate A$\Rightarrow$ BCE and AE, AB and A are removed from L_{ABCE}. Then, we have $\frac{|ABCE|}{|B|} = \frac{1}{2}$. So, we generate the generic rule B$\Rightarrow$ACE and we delete B from L_{ABCE}. From, the evaluation of $\frac{|ABCE|}{|E|} = \frac{1}{2}$, we have to generate E$\Rightarrow$ ABC, and to remove E from L_{ABCE}, which is found to be empty.

For I = BCE we have |BCE| $\geq$ minconf, then we generate the generic rule $\emptyset \Rightarrow$BCE. Next, we have to apply the same process, respectively, for I = AC, I = BE and I = C. The resulting generic base is depicted by Table 2.

Obviously, the introduced $\mathcal{IGB}$ generic base is largely more compact than that proposed by Bastide *et al.* (8 generic rules vs 17). Comparatively to that proposed by *Phan*, the $\mathcal{IGB}$ generic base is not more compact but it presents the informative property. The IGB algorithm is currently under implementation, and we have to assess its compactness output versus those presented in the literature survey.

Algorithm 1.2: IGB

Data:

1. $\mathcal{FCI}$: set of frequent closed itemsets and their associated supports.
2. minconf

Result: $\mathcal{IGB}$:Informative generic base
begin
  foreach *frequent closed itemset* $I \in \mathcal{FCI}$ / $I \neq \emptyset$ **do**
    if *(*$|I| \geq$ *minconf)* **then**
      R= $\emptyset \Rightarrow$ I
      R.supp=|I|
      R.conf=|I|
      $\mathcal{IGB}$=$\mathcal{IGB} \cup$ R
    else
      L_{I_1}={non empty minimal generators of frequent closed itemset I_1 | $I_1 \subseteq$ I}
      foreach *minimal generator* $g \in L_{I_1}$ **do**
        if *(*$\frac{|I|}{|g|} \geq$ *minconf* $\wedge$ $I \neq g$*)* **then**
          R= g $\Rightarrow$I-g
          R.supp=|I|
          R.conf=$\frac{|I|}{|g|}$
          $\mathcal{IGB}$=$\mathcal{IGB} \cup$ R
          L_{I_1}=L_{I_1}-$\{g' | g \subset g'\}$
        L_{I_1}=L_{I_1}-g
end

4.2 Association rule derivation

In this subsection, we will try to axiomatize the derivation of association rules from the the $\mathcal{IGB}$ generic base. In the following, we provide a set of inference axioms and we prove their soundness. Then, we show that the $\mathcal{IGB}$ generic base is informative.

Proposition 2. *Let* $\mathcal{IGB}_\mathcal{K}$ *and* $\mathcal{AR}_\mathcal{K}$ *be a generic base and the set of valid association rules, respectively, that can be drawn from a context* $\mathcal{K}$*. Then, the following inference axioms are sound:*

A0.Conditional Reflexivity: *If* $X \stackrel{c}{\Rightarrow} Y \in IGB_\mathcal{K} \wedge X \neq \emptyset$ *then* $X \stackrel{c}{\Rightarrow} Y \in \mathcal{AR}_\mathcal{K}$

A1.Conditional Decomposition: *If* $X \stackrel{c}{\Rightarrow} Y \in IGB_\mathcal{K} \wedge X \neq \emptyset$ *then* $X \stackrel{c'}{\Rightarrow} Z \in \mathcal{AR}_\mathcal{K} / Z \subset Y \wedge c' = \frac{|XZ|}{|X|}$.

A2.Augmentation *If* $X \stackrel{c}{\Rightarrow} Y \in IGB_\mathcal{K}$ *then* $X \cup Z \stackrel{c'}{\Rightarrow} Y$-$\{Z\} \in \mathcal{AR}_\mathcal{K}$ $/Z \subset Y \wedge c' = \frac{|Y|}{|XZ|}$.

Proof. **A0.Conditional Reflexivity:** follows from the proper definition of the $\mathcal{IGB}$ generic base. The condition $X \neq \emptyset$ ensures the non emptiness of the derived rule.

Implication	Support	Confidence
C⇒ABE	$\frac{2}{5}$	$\frac{1}{2}$
A⇒BCE	$\frac{2}{5}$	$\frac{2}{3}$
B⇒ACE	$\frac{2}{5}$	$\frac{1}{2}$
E⇒ABC	$\frac{2}{5}$	$\frac{1}{2}$
$\emptyset$ ⇒BCE	$\frac{3}{5}$	$\frac{3}{5}$
$\emptyset$ ⇒AC	$\frac{3}{5}$	$\frac{3}{5}$
$\emptyset$ ⇒BE	$\frac{4}{5}$	$\frac{4}{5}$
$\emptyset$ ⇒C	$\frac{4}{5}$	$\frac{4}{5}$

Table 2. The $\mathcal{IGB}$ generic base.

A1.Conditional Decomposition: Since, X$\stackrel{c}{\Rightarrow}$Y $\in IGB_{\mathcal{K}}$ then conf(X$\stackrel{c}{\Rightarrow}$Y)=c $\geq$ minconf. Since Z $\subset$ XY, we have |XZ|$\geq$ |XY| and then $\frac{|XZ|}{|X|} \geq \frac{|XY|}{|X|} \geq$ minconf. Hence, X$\stackrel{c'}{\Rightarrow}$Z is a valid rule with a confidence value c$'$=$\frac{|XZ|}{|X|}$.The condition $X \neq \emptyset$ ensures the non emptiness of the derived rule.

A2.Augmentation Since, X$\stackrel{c}{\Rightarrow}$Y $\in IGB_{\mathcal{K}}$ then conf(X$\stackrel{c}{\Rightarrow}$Y)=c $\Leftrightarrow$ $\frac{|XY|}{|X|}$=c $\geq$ minconf. From X $\subset$ XZ, we have |X| >|XZ| > and minconf < $\frac{|XY|}{|X|} < \frac{|XYZ|}{|XZ|}$. Hence, X $\cup$ Z$\stackrel{c'}{\Rightarrow}$Y-{Z} is a valid rule, with a confidence value c$'$= $\frac{|Y|}{|XZ|}$.

Proposition 3. *The $\mathcal{IGB}$ generic base is informative.*

Proof. To prove that the $\mathcal{IGB}$ generic base is informative, it is sufficient to show that it contains all the necessary information to determine the support of an itemset in a derived rule. Therefore, it means that we have to be able to reconstitute all closed itemset by concatenation of the premise and the conclusion of a generic rule [6]. The algorithm considers all the discovered frequent closed itemsets. Hence for a given frequent closed itemset, say c, it tries to find the smallest minimal generator, say g_s, associated to frequent closed itemsets subsumed by c and fulfilling the minsup constraint. Therefore, the algorithm generates the following generic rule $g_s \Rightarrow$c. Since $g_s \subset$ c, then $g_s \cup$ c=c. Then by construction, all frequent closed itemsets can be reconstituted from the $\mathcal{IGB}$ generic base.

Conclusion: The $\mathcal{IGB}$ generic base is informative.

In what follows, we present the AR-DERIV algorithm, whose pseudo-code is given by Algorithm1.3, permitting to derive all valid association rules from the $\mathcal{IGB}$ generic base.

5 Conclusion

In this paper, we presented a critical survey of the reported approaches for defining generic bases of association rules. Then, we introduced a novel generic

[6] It is known that the support of an itemset is equal to the support of the smallest closed itemset containing it.

Algorithm 1.3: A-R-DERIV

Data: $\mathcal{IGB}_{\mathcal{K}}$: Informative generic base
Result: $\mathcal{AR}_{\mathcal{K}}$:set of valid association rules
begin
  foreach $R{:}X \stackrel{c}{\Rightarrow} Y \in \mathcal{IGB}_{\mathcal{K}}$ **do**
    /* Applying Reflexivity axiom*/
    if $X \neq \emptyset$ **then** $\mathcal{AR}_{\mathcal{K}}$=$\mathcal{AR}_{\mathcal{K}} \cup$ X$\stackrel{c}{\Rightarrow} Y$
    if $|Y| > 1$ **then**
      foreach $Z \mid Z \subset Y$ **do**
        –
        /*Applying Decomposition axiom*/
        R'=X⇒Z
        s=GET-SMALLEST-SUPPORT(XZ, $\mathcal{IGB}_{\mathcal{K}}$)
        /* the GET-SMALLEST-SUPPORT function yields the support value of the smallest closed itemset containing XZ*/
        c'= $\frac{c \times s}{|XY|}$
        $\mathcal{AR}_{\mathcal{K}}$=$\mathcal{AR}_{\mathcal{K}} \cup$ X$\stackrel{c'}{\Rightarrow} Z$
        /*Applying Augmentation axiom*/
        R'=XZ⇒Y-{Z}
        s=GET-SMALLEST-SUPPORT(XZ,$\mathcal{IGB}_{\mathcal{K}}$)
        c'=$\frac{|XY|}{s}$
        $\mathcal{AR}_{\mathcal{K}}$=$\mathcal{AR}_{\mathcal{K}} \cup$ XZ$\stackrel{c'}{\Rightarrow} Y - \{Z\}$
end

base, which is sound and informative. We also provided a set of sound inference axioms for deriving all association rules from the introduced generic base of association rules. The reported algorithms are currently under implementation in order to include them in a Information Retrieval prototype. Specially, we are interested in assessing the well-known IR metrics, namely recall and precision, by using the introduced generic rule base in a query expansion process.

References

1. Bastide, Y., Pasquier, N., Taouil, R., Lakhal, L., Stumme, G.: Mining minimal non-redundant association rules using frequent closed itemsets. In: Proceedings of the Intl. Conference DOOD'2000, LNCS, Springer-verlag. (2000) 972–986
2. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Intelligent structuring and reducing of association rules with formal concept analysis. In: Proc. KI'2001 conference, LNAI 2174, Springer-verlag. (2001) 335–350
3. Zaki, M.J.: Generating non-redundant association rules. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,Boston, MA. (2000) 34–43
4. Ganter, B., Wille, R.: Formal Concept Analysis. Springer-Verlag (1999)
5. BenYahia, S., Cherif, C.L., Mineau, G., Jaoua, A.: Découverte des règles associatives non redondantes : application aux corpus textuels. Revue d'Intelligence

Artificielle (special issue of Intl. Conference of Journées francophones d'Extraction et Gestion des Connaissances(EGC'2003), Lyon, France **17** (2003) 131–143
6. Barbut, M., Monjardet, B.: Ordre et classification. Algèbre et Combinatoire. Hachette, Tome II (1970)
7. BenYahia, S., Nguifo, E.M.: Revisiting generic bases of association rules. In: Proceedings of 6th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2004) (Springer-Verlag) to appear, Zaragoza, Spain. (2004)
8. Guigues, J., Duquenne, V.: Familles minimales d'implications informatives résultant d'un tableau de données binaires. Mathématiques et Sciences Humaines (1986) 5–18
9. Luxenburger, M.: Implication partielles dans un contexte. Mathématiques et Sciences Humaines **29** (1991) 35–55
10. Kryszkiewicz, M.: Concise representations of association rules. In Hand, D.J., Adams, N., Bolton, R., eds.: Proceedings of Pattern Detection and Discovery, ESF Exploratory Workshop, London, UK. Volume 2447 of Lecture Notes in Computer Science., Springer (2002) 92–109
11. Luong, V.P.: Raisonnement sur les règles d'association. In: Proceedings 17ème Journées Bases de Données Avancées BDA'2001, Agadir (Maroc), Cépaduès Edition. (2001) 299–310
12. Armstrong, W.: Dependency structures of database relationships. In: IFIP Congress. (1974) 580–583