Mining Minimal Non-Redundant Association Rules using Frequent Closed Itemsets

Yves Bastide[1], Nicolas Pasquier[1], Rafik Taouil[1], Gerd Stumme[1,2], Lotfi Lakhal[1]

[1] L.I.M.O.S., Université Blaise Pascal – Clermont-Ferrand II,
Complexe des Cézeaux, 24 Avenue des Landais, F–63177 Aubière cedex, France;
{bastide,pasquier,taouil,lakhal}@libd2.univ-bpclermont.fr
[2] Technische Universität Darmstadt, Fachbereich Mathematik, Schloßgartenstr. 7,
D–64289 Darmstadt, Germany; stumme@mathematik.tu-darmstadt.de

Abstract. The problem of the relevance and the usefulness of extracted association rules is of primary importance because, in the majority of cases, real-life databases lead to several thousands association rules with high confidence and among which are many redundancies. Using the closure of the Galois connection, we define two new bases for association rules which union is a generating set for all valid association rules with support and confidence. These bases are characterized using frequent closed itemsets and their generators; they consist of the non-redundant exact and approximate association rules having minimal antecedents and maximal consequents, i.e. the most relevant association rules. Algorithms for extracting these bases are presented and results of experiments carried out on real-life databases show that the proposed bases are useful, and that their generation is not time consuming.

1 Introduction

The purpose of association rule extraction, introduced in [AIS93], is to discover significant relations between binary attributes extracted from databases. An example of association rule extracted from a database of supermarket sales is: "cereals $\wedge$ sugar $\rightarrow$ milk (support 7%, confidence 50%)". This rule states that the customers who buy cereals and sugar also tend to buy milk. The *support* defines the range of the rule, i.e. the proportion of customers who bought the three items among all customers, and the *confidence* defines the precision of the rule, i.e. the proportion of customers who bought milk among those who bought cereals and sugar. An association rule is considered relevant for decision making if it has support and confidence at least equal to some minimal support and confidence thresholds, *minsupport* and *minconfidence*, defined by the user.
The problem of relevance and usefulness of the result is related to the number of extracted association rules – that is in general very large – and to the presence of a huge proportion of redundant rules, i.e. rules conveying the same information, among them. Even though the visualization of a relatively significant number of rules can be simplified by the use of visualization tools such as the Rule Visualizer system [KMR+94], suppressing redundant association rules requires

other solutions. Moreover, as the redundant association rules represent the majority of the extracted rules for several kinds of data, their suppression reduces considerably the number of rules to be managed during the visualization.

Example 1. In order to illustrate the problem of redundant association rules, nine association rules extracted from UCI KDD's archives's dataset MUSHROOMS[1] describing the characteristics of 8 416 mushrooms are presented below. These nine rules have identical supports and confidences of 51% and 54% respectively, and the item "free gills" in the antecedent:

1) free gills → eatable
2) free gills → eatable, partial veil
3) free gills → eatable, white veil
4) free gills, white veil → eatable
5) free gills, partial veil → eatable
6) free gills → eatable, partial veil, white veil
7) free gills, partial veil → eatable, white veil
8) free gills, white veil → eatable, partial veil
9) free gills, partial veil, white veil → eatable

Obviously, given rule 6, rules 1 to 5 and 7 to 9 are redundant, since they do not convey any additional information to the user. Rule 6 has minimal antecedent and maximal consequent and it is the most informative among these nine rules. In order to improve the relevance and the usefulness of extracted rules, only rule 6 should be extracted and presented to the user.

Several methods have been proposed in the literature to reduce the number of extracted association rules. Generalized association rules [HF95,SA95] are defined using a taxonomy of the items; they are rules between sets of items that belong to different levels of the taxonomy. The use of statistic measures other than confidence such as conviction, Pearson's correlation or χ^2 test is studied in [BMS97,SBM98]. In [Hec96,PSM94,ST96], the use of deviation measures, i.e. measures of distance between association rules, defined according to their supports and confidences, is proposed. In [BAG99,NLHP98,SVA97], the use of item constraints, that are boolean expressions defined by the user, in order to specify the form of the association rules that will be presented to the user is proposed. The approach proposed in [BG99] is to present to the user rules with maximal antecedents, called A-maximal rules, that are rules for which the population of objects concerned is reduced when an item is added to the antecedent. In [PBTL99c], we adapt the Duquenne-Guigues basis for global implications [DG86,GW99] and the proper basis for partial implications [Lux91] to the association rules framework. It is demonstrated that these bases are minimal with respect to the number of extracted association rules. However, none of these methods allows to generate the non-redundant association rules with minimal antecedents and maximal consequents which we believe are the most relevant and useful from the point of view of the user.

1.1 Contribution

In the rest of the paper, two kinds of association rules are distinguished:

[1] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/mushroom/

- Exact association rules whose confidence is equal to 100%, i.e. which are valid for all the objects of the context. These rules are written $l \Rightarrow l'$.
- Approximate association rules whose confidence is lower than 100%, i.e. which are valid for a proportion of objects of the context equal to their confidence. These rules are written $l \rightarrow l'$.

The solution proposed in this paper consists in generating *bases*, or *reduced covers*, for association rules. These bases contain no redundant rule, being thus of smaller size. Our goal is to limit the extraction to the most informative association rules from the point of view of the user.
Using the semantic for the extraction of association rules based on the closure of the Galois connection [PBTL98], the *generic basis for exact association rules* and the *informative basis for approximate association rules* are defined. They are constructed using the frequent closed itemsets and their generators, and they minimize the number of association rules generated while maximizing the quantity and the quality of the information conveyed. They allow for:

1. The generation of only the most informative non-redundant association rules, i.e. of the most useful and relevant rules: those having a minimal antecedent (left-hand side) and a maximal consequent (right-hand side). Thus redundant rules which represent in certain databases the majority of extracted rules, particularly in the case of dense or correlated data for which the total number of valid rules is very large, will be pruned.
2. The presentation to the user of a set of rules covering all the attributes of the database, i.e. containing rules where the union of the antecedents (resp. consequents) is equal to the unions of the antecedents (resp. consequents) of all the association rules valid in the context. This is necessary in order to discover rules that are "surprising" to the user, which constitute important information that it is necessary to consider [Hec96,PSM94,ST96].
3. The extraction of a set of rules without any loss of information, i.e. conveying all the information conveyed by the set of all valid association rules. It is possible to deduce efficiently, without access to the dataset, all valid association rules with their supports and confidences from these bases.

The union of these two bases thus constitutes a small non-redundant generating set for all valid association rules, their supports and their confidences.
In section 2, we recall the semantic for association rules based on the Galois connection. The new bases we propose and algorithms for generating them are defined in section 3. Results of experiments we conducted on real-life datasets are presented in section 4 and section 5 concludes the paper.

2 Semantic for association rules based on the Galois connection

The association rule extraction is performed from a data mining context, that is a triplet $\mathcal{D} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, where $\mathcal{O}$ and $\mathcal{I}$ are finite sets of objects and items respectively, and $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$ is a binary relation. Each couple $(o, i) \in \mathcal{R}$ denotes the fact that the object $o \in \mathcal{O}$ is related to the item $i \in \mathcal{I}$.

Example 2. A data mining context $\mathcal{D}$ constituted of six objects (each one identified by its *OID*) and five items is represented in the table 1. This context is used as support for the examples in the rest of the paper.

OID	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

Table 1. Data mining context $\mathcal{D}$.

The closure operator γ of the Galois connection [GW99] is the composition of the application ϕ, that associates with $O \subseteq \mathcal{O}$ the items common to all objects $o \in O$, and the application ψ, that associates with an itemset $l \subseteq \mathcal{I}$ the objects related to all items $i \in l$ (the objects "containing" l).

Definition 1 (Frequent itemsets). *A set of items $l \subseteq \mathcal{I}$ is called an* itemset. *The support of an itemset l is the percentage of objects in $\mathcal{D}$ containing l: $support(l) = |\psi(l)| \,/\, |\mathcal{O}|$. l is a frequent itemset if $support(l) \geq$* minsupport.

Definition 2 (Association rules). *An association rule r is an implication between two frequent itemsets $l_1, l_2 \subseteq \mathcal{I}$ of the form $l_1 \rightarrow (l_2 \setminus l_1)$ where $l_1 \subset l_2$. The support and the confidence of r are defined as: $support(r) = support(l_2)$ and $confidence(r) = support(l_2) \,/\, support(l_1)$.*

The closure operator $\gamma = \phi \circ \psi$ associates with an itemset l the maximal set of items common to all the objects containing l, i.e. the intersection of these objects. Using this closure operator, we define the frequent closed itemsets that constitute a minimal non-redundant generating set for all frequent itemsets and their supports, and thus for all association rules, their supports and their confidences. This property comes from the facts that the support of a frequent itemset is equal to the support of its closure and that the maximal frequent itemsets are maximal frequent closed itemsets [PBTL98].

Definition 3 (Frequent closed itemsets). *A frequent itemset $l \subseteq \mathcal{I}$ is a frequent closed itemset iff $\gamma(l) = l$. The smallest (minimal) closed itemset containing an itemset l is $\gamma(l)$, i.e. the closure of l.*

In order to extract the frequent closed itemsets, the Close [PBTL98,PBTL99a] and the A-Close [PBTL99b] algorithms perform a breadth-first search for the *generators* of the frequent closed itemsets in a levelwise manner.

Definition 4 (Generators). *An itemset $g \subseteq \mathcal{I}$ is a (minimal) generator of a closed itemset l iff $\gamma(g) = l$ and $\nexists g' \subseteq \mathcal{I}$ with $g' \subset g$ such that $\gamma(g') = l$. A generator of cardinality k is called a k-generator.*

2.1 Extracting frequent closed itemsets and their generators with the Close algorithm

The Close algorithm is an iterative algorithm for the extraction of all frequent closed itemsets. It courses generators of the frequent closed itemsets in a levelwise manner. During the k^{th} iteration of the algorithm, a set FCC_k of candidates is considered. Each element of this set consists of three fields: a candidate k-generator, its closure (which is a candidate closed itemset), and their support (the supports of the generator and its closure being identical). At the end of the k^{th} iteration, the algorithm stores a set FC_k containing the frequent k-generators, their closures which are frequent closed itemsets, and their supports.
The algorithm starts by initializing the set FCC_1 of the candidate 1-generators with the list of the 1-itemsets of the context and then carries out some iterations. During each iteration k:

1. The closures of all k-generators and their supports are computed. This computation is based on the property that the closure of an itemset is equal to the intersection of all the objects in the context containing it. The number of these objects provides the support of the generator. Only one scan of the context is thus necessary to determine the closures and the supports of all the k-generators.
2. All frequent k-generators, which support is greater or equal to *minsupport*, their closures and their supports are inserted in the set FC_k of frequent closed itemsets identified during the iteration k.
3. The set of candidate $(k+1)$-generators (used during the following iteration) is constructed, by joining the frequent k-generators in the set FC_k as follows.
   (a) The candidate $(k+1)$-generators are created by joining the k-generators in FC_k that have the same $k-1$ first items. For instance, the 3-generators $\{ABC\}$ and $\{ABD\}$ will be joined in order to create the candidate 4-generator $\{ABCD\}$.
   (b) The candidate $(k+1)$-generators that are known to be either infrequent or non-minimal, because one of their subset is either infrequent or non-minimal, are then removed. These generators are identified by the absence of at least one their subsets of size k among the frequent k-generators of FC_k.
   (c) A third phase removes among the remaining generators those which closures were already computed. Such a $(k+1)$-generator g is easily identified since it is included in the closure of a frequent k-generator g' in FC_k: $g' \subset g \subset \gamma(g')$ (i.e. it is not a minimal generator).

The algorithm stops when no new candidate generator can be created. The A-Close algorithm, developed in order to improve the effectiveness of the extraction in the case of slightly correlated data, does not compute the closures of the candidate generators during the iterations, but during an ultimate scan carried out after the end of these iterations.

Example 3. Figure 1 shows the execution of the Close algorithm on the context $\mathcal{D}$ for a minimal support threshold of 2/6. The algorithm carries out two iterations, and thus two dataset scans.

Scan $\mathcal{D}$ $\longrightarrow$

FCC_1

Generator	Closed itemset	Support
{A}	{AC}	3/6
{B}	{BE}	5/6
{C}	{C}	5/6
{D}	{ACD}	1/6
{E}	{BE}	5/6

Suppressing infrequent itemsets $\longrightarrow$

FC_1

Generator	Closed itemset	Support
{A}	{AC}	3/6
{B}	{BE}	5/6
{C}	{C}	5/6
{E}	{BE}	5/6

Scan $\mathcal{D}$ $\longrightarrow$

FCC_2

Generator	Closed itemset	Support
{AB}	{ABCE}	2/6
{AE}	{ABCE}	2/6
{BC}	{BCE}	4/6
{CE}	{BCE}	4/6

Suppressing infrequent itemsets $\longrightarrow$

FC_2

Generator	Closed itemset	Support
{AB}	{ABCE}	2/6
{AE}	{ABCE}	2/6
{BC}	{BCE}	4/6
{CE}	{BCE}	4/6

Fig. 1. Extracting frequent closed itemsets from $\mathcal{D}$ with Close for $minsupport = 2/6$.

Experimental results showed that these algorithms are particularly efficient for mining association rules from dense or correlated data that represent an important part of real life databases.

3 Minimal non-redundant association rules

As pointed out in example 1, it is desirable that only the non-redundant association rules with minimal antecedent and maximal consequent, i.e. the most useful and relevant rules, are extracted and presented to the user. Such rules are called *minimal non-redundant association rules*.

Support and confidence indicate the range and the precision of the rule, and thus, must be taken into account for characterizing the redundant association rules. In previous works concerning the reduction of redundant implication rules (functional dependancies), such as the definition of the canonical cover [BB79,Mai80], the notion of non-redundancy considered is related to the inference system using Armstrong axioms [Arm74]. This notion is not to be confused with the notion of non-redundancy we consider here. To our knowledge, such an inference system for association rules, that takes into account supports and confidences of the rules, does not exist. The principle of minimal non-redundant association rules as defined hereafter is to identify the most informative association rules considering the fact that in practice, the user cannot infer all other valid rules from the rules extracted while visualizing them.

An association rule is redundant if it conveys the same information – or less general information – than the information conveyed by another rule of the same usefulness and the same relevance. An association rule $r \in E$ is non-redundant and minimal if there is no other association rule $r' \in E$ having the same support and the same confidence, of which the antecedent is a subset of the antecedent of r and the consequent is a superset of the consequent of r.

Definition 5 (Minimal non-redundant association rules). *An association rule $r : l_1 \rightarrow l_2$ is a minimal non-redundant association rule iff there does not exist an association rule $r' : l'_1 \rightarrow l'_2$ with* support(r) = support(r'), confidence(r) = confidence(r'), *$l'_1 \subseteq l_1$ and $l_2 \subseteq l'_2$.*

Based on this definition, we characterize the generic basis for exact association rules and the informative basis for approximate association rules, constituted of the minimal non-redundant exact and approximate association rules respectively.

3.1 Generic basis for exact association rules

The exact association rules, of the form $r : l_1 \Rightarrow (l_2 \setminus l_1)$, are rules between two frequent itemsets l_1 and l_2 whose closures are identical: $\gamma(l_1) = \gamma(l_2)$. Indeed, from $\gamma(l_1) = \gamma(l_2)$ we deduce that $l_1 \subset l_2$ and $support(l_1) = support(l_2)$, and thus $confidence(r) = 1$. Since the maximum itemset among these itemsets (which have same supports) is the itemset $\gamma(l_2)$, all supersets of l_1 that are subsets of $\gamma(l_2)$ have the same support, and the rules between two of these itemsets are exact rules.
Let $G_{\gamma(l_2)}$ be the set of generators of the frequent closed itemset $\gamma(l_2)$. By definition, the minimal itemsets that are supersets of l_1 and are subsets of $\gamma(l_2)$ are the generators $g \in G_{\gamma(l_2)}$. We thus conclude that rules of the form $g \Rightarrow (\gamma(l_2) \setminus g)$ between generators $g \in G_{\gamma(l_2)}$ and the frequent closed itemset $\gamma(l_2)$ are the rules of minimal antecedents and maximal consequents among the rules between the supersets of l_1 and the subsets of $\gamma(l_2)$. The generalization of this property to the set of frequent closed itemsets defines the generic basis consisting of all non-redundant exact association rules with minimal antecedents and maximal consequents, as characterized in definition 5.

Definition 6 (Generic basis for exact association rules). *Let FC be the set of frequent closed itemsets extracted from the context and, for each frequent closed itemset f, let denote G_f the set of generators of f. The generic basis for exact association rules is:*

$$GB = \{r : g \Rightarrow (f \setminus g) \mid f \in FC \ \wedge \ g \in G_f \ \wedge \ g \neq f\}.$$

The condition $g \neq f$ ensures that rules of the form $g \Rightarrow \varnothing$ that are non-informative are discarded. The following proposition states that the generic basis does not lead to any loss of information.

Proposition 1. *(i) All valid exact association rules, their supports and their confidences (that are equals to 100%) can be deduced from the rules of the generic basis and theirs supports. (ii) The generic basis for exact association rules contains only minimal non redundant-rules.*

Proof. Let $r : l_1 \Rightarrow (l_2 \setminus l_1)$ be a valid exact association rule between two frequent itemsets with $l_1 \subset l_2$. Since $confidence(r) = 100\%$ we have $support(l_1) = support(l_2)$. Given the property that the support of an itemset is equal to the support of its closure, we deduce that $support(\gamma(l_1)) = support(\gamma(l_2)) \Longrightarrow \gamma(l_1) =$

$\gamma(l_2) = f$. The itemset f is a frequent closed itemset $f \in FC$ and, obviously, there exists a rule $r' : g \Rightarrow (f \setminus g) \in GB$ such that g is a generator of f for which $g \subseteq l_1$ and $g \subset l_2$. We show that the rule r and its support can be deduced from the rule r' and its support. Since $g \subseteq l_1 \subset l_2 \subseteq f$, the rule r can be derived from the rule r'. From $\gamma(l_1) = \gamma(l_2) = f$, we deduce that $support(r) = support(l_2) = support(\gamma(l_2)) = support(f) = support(r')$. $\square$

Algorithm for constructing the generic basis

The pseudo-code of the Gen-GB algorithm for constructing the generic basis for exact association rules using the frequent closed itemsets and their generators is presented in algorithm 1. Each element of a set FC_k consists of three elements: *generator*, *closure* and *support*.

Algorithm 1 Constructing the generic basis with Gen-GB.

Input : sets FC_k of k-groups of frequent k-generators;
Output : set GB of exact association rules of the generic basis;
1) $GB \leftarrow \{\}$
2) **forall** set $FC_k \in FC$ **do begin**
3)   **forall** k-generator $g \in FC_k$ **such that** $g \neq \gamma(g)$ **do begin**
4) $GB \leftarrow GB \cup \{(r : g \Rightarrow (\gamma(g) \setminus g), \gamma(g).support)\}$;
5)   **end**
6) **end**
7) **return** GB;

The algorithm starts by initializing the set GB with the empty set (step 1). Each set FC_k of frequent k-groups is then examined successively (steps 2 to 6). For each k-generator $g \in FC_k$ of the frequent closed itemset $\gamma(g)$ for which g is different from its closure $\gamma(g)$ (steps 3 to 5), the rule $r : g \Rightarrow (\gamma(g) \setminus g)$, whose support is equal to the support of g and $\gamma(g)$, is inserted into GB (step 4). The algorithm returns finally the set GB containing all minimal non-redundant exact association rules between generators and their closures (step 7).

Example 4. The generic basis for exact association rules extracted from the context $\mathcal{D}$ for a minimal support threshold of 2/6 is presented in Table 2. It contains seven rules whereas fourteen exact association rules are valid on the whole.

3.2 Informative basis for approximate association rules

Each approximate association rule $l_1 \rightarrow (l_2 \setminus l_1)$, is a rule between two frequent itemsets l_1 and l_2 such that the closure of l_1 is a subset of the closure of l_2: $\gamma(l_1) \subset \gamma(l_2)$. The non-redundant approximate association rules with minimal antecedent l_1 and maximal consequent $(l_2 \setminus l_1)$ are deduced from this characterisation.
Let f_1 be the frequent closed itemset which is the closure of l_1, and g_1 a generator of f_1 such as $g_1 \subseteq l_1 \subseteq f_1$. Let f_2 be the frequent closed itemset which is the closure of l_2 and g_2 a generator of f_2 such as $g_2 \subseteq l_2 \subseteq f_2$. The rule $g_1 \Rightarrow (f_2 \setminus g_1)$

Generator	Closure	Exact rule	Support
{A}	{AC}	A ⇒ C	3/6
{B}	{BE}	B ⇒ E	5/6
{C}	{C}		
{E}	{BE}	E ⇒ B	5/6
{AB}	{ABCE}	AB ⇒ CE	2/6
{AE}	{ABCE}	AE ⇒ BC	2/6
{BC}	{BCE}	BC ⇒ E	4/6
{CE}	{BCE}	CE ⇒ B	4/6

Table 2. Generic basis for exact association rules extracted from $\mathcal{D}$ for *minsupport* $= 2/6$.

between the generator g_1 and the frequent closed itemset f_2 is the minimal non-redundant rule among the rules between an itemset of the interval[2] $[g_1, f_1]$ and an itemset of the interval $[g_2, f_2]$. Indeed, the generator g_1 is the minimal itemset whose closure is f_1, which means that the antecedent g_1 is minimal and that the consequent $(f_2 \setminus g_1)$ is maximal since f_2 is the maximal itemset of the interval $[g_2, f_2]$. The generalization of this property to the set of all rules between two itemsets l_1 and l_2 defines the informative basis which thus consists of all the non-redundant approximate association rules of minimal antecedents and maximal consequents characterized in definition 5.

Definition 7 (Informative basis for approximate association rules). *Let FC be the set of frequent closed itemsets and let denote G the set of their generators extracted from the context. The informative basis for approximate association rules is:*

$$IB = \{r : g \to (f \setminus g) \mid f \in FC \ \wedge \ g \in G \ \wedge \ \gamma(g) \subset f\}.$$

Proposition 2. *(i) All valid approximate association rules, their supports and confidences, can be deduced from the rules of the informative basis, their supports and theirs confidences. (ii) All rules in the informative basis re minimal non-redundant approximate association rules.*

Proof. Let $r : l_1 \to (l_2 \setminus l_1)$ be a valid approximate association rule between two frequent itemsets with $l_1 \subset l_2$. Since $confidence(r) < 1$ we also have $\gamma(l_1) \subset \gamma(l_2)$. For any frequent itemsets l_1 and l_2, there is a generator g_1 such that $g_1 \subset l_1 \subseteq \gamma(l_1) = \gamma(g_1)$ and a generator g_2 such that $g_2 \subset l_2 \subseteq \gamma(l_2) = \gamma(g_2)$. Since $l_1 \subset l_2$, we have $l_1 \subseteq \gamma(g_1) \subset l_2 \subseteq \gamma(g_2)$ and the rule $r' : g_1 \to (\gamma(g_2) \setminus g_1)$ belongs to the informative basis IB. We show that the rule r, its support and its confidence can be deduced from the rule r', its support and its confidence. Since $g_1 \subset l_1 \subseteq \gamma(g_1) \subset g_2 \subset l_2 \subseteq \gamma(g_2)$, the antecedent and the consequent of r can be rebuilt starting from the rule r'. Moreover, we have $\gamma(l_2) = \gamma(g_2)$ and thus $support(r) = support(l_2) = support(\gamma(g_2)) = support(r')$. Since $g_1 \subset l_1 \subseteq \gamma(g_1)$, we have $support(g_1) = support(l_1)$ and we thus deduce that: $confidence(r) = support(l_2) \,/\, support(l_1) = support(\gamma(g_2)) \,/\, support(g_1) = confidence(r')$. □

[2] The interval $[l_1, l_2]$ contains all the supersets of l_1 that are subsets of l_2.

From the definition of the informative basis we deduce the definition of the transitive reduction of the informative basis that is itself a basis for all approximate association rules. We note $l_1 \lessdot l_2$ if the itemset l_1 is an immediate predecessor of the itemset l_2, i.e. $\nexists l_3$ such that $l_1 \subset l_3 \subset l_2$. The transitive rules of the informative basis are of the form $r : g \rightarrow (f \setminus g)$ for a frequent closed itemset f and a frequent generator g such that $\gamma(g) \subset f$ and $\gamma(g)$ is not an immediate predecessor of $f in FC$: $\gamma(g) \not\lessdot f$. The transitive reduction of the informative basis thus contains the rules with the form $r : g \rightarrow (f \setminus g)$ for a frequent closed itemset f and a frequent generator g such as $\gamma(g) \lessdot f$.

Definition 8 (Transitive reduction of the informative basis). *Let FC be the set of frequent closed itemsets and let denote G the set of their generators extracted from the context. The transitive reduction of the informative basis for approximate association rules is:*

$$RI = \{r : g \rightarrow (f \setminus g) \mid f \in FC \ \wedge \ g \in G \ \wedge \ \gamma(g) \lessdot f\}.$$

Obviously, it is possible to deduce all the association rules of the informative basis with their supports and their confidences, and thus all the valid approximate rules, from the rules of this transitive reduction, their supports and their confidences. This reduction makes it possible to decrease the number of approximate rules extracted by preserving the rules which confidences are the highest (since the transitive rules have confidences lower than the non-transitive rules by construction) without losing any information.

Constructing the transitive reduction of the informative basis

The pseudo code of the Gen-RI algorithm for constructing the transitive reduction of the informative basis for the approximate association rules using the set of frequent closed itemsets and their generators is presented in algorithm 2. Each element of a set FC_k consists of three fields: *generator*, *closure* and *support*. The algorithm constructs for each generator g considered a set $Succ_g$ containing the frequent closed itemsets that are immediate successors of the closure of g.
The algorithm starts by initializing the set RI with the empty set (step 1). Each set FC_k of frequent k-groups is then examined successively in the increasing order of the values of k (steps 2 to 14). For each k-generator $g \in FC_k$ of the frequent closed itemset $\gamma(g)$ (steps 3 to 18), the set $Succ_g$ of the successors of the closure of $\gamma(g)$ is initialized with the empty set (step 4) and the sets S_j of frequent closed j-itemsets that are supersets of $\gamma(g)$ for $|\gamma(g)| < j \leq \mu$[3] are constructed (steps 5 to 7). The sets S_j are then considered in the ascending order of the values of j (steps 8 to 17). For each itemset $f \in S_j$ that is not a superset of an immediate successor of $\gamma(g)$ in $Succ_g$ (step 10), f is inserted in $Succ_g$ (step 11) and the confidence of the rule $r : g \rightarrow (f \setminus g)$ is computed (step 12). If the confidence of r is greater or equal to the minimal confidence threshold *minconfidence*, the rule r is inserted in RI (steps 13 to 15). When all the generators of size lower than μ have been considered, the algorithm returns the set RI (step 20).

[3] We denote μ the size of the longest maximal frequent closed itemsets.

Algorithm 2 Generating the transitive reduction of the informative basis with Gen-RI.

Input : sets FC_k of k-groups of frequent k-generators; *minconfidence* threshold;
Output : Transitive reduction of the informative basis RI;

$RI \leftarrow \{\}$
for $(k \leftarrow 1;\ k \leq \mu\text{-}1;\ k{+}{+})$ **do begin**
    forall k-generator $g \in FC_k$ **do begin**
        $Succ_g \leftarrow \{\}$;
        for $(j = |\gamma(g)|;\ j \leq \mu;\ j{+}{+})$ **do begin**
            $S_j \leftarrow \{f \in FC \mid f \supset \gamma(g) \ \wedge\ |f| = j\}$;
        end
        for $(j = |\gamma(g)|;\ j \leq \mu;\ j{+}{+})$ **do begin**
            forall frequent closed itemset $f \in S_j$ **do begin**
                if $(\nexists s \in Succ_g \mid s \subset f)$ **then do begin**
                    $Succ_g \leftarrow Succ_g \cup f$;
                    $r.confidence \leftarrow f.support/g.support$;
                    if $(r.confidence \geq minconfidence)$
                    then $RI \leftarrow RI \cup \{r : g \rightarrow (f \setminus g), r.confidence, f.support\}$;
                endif
            end
        end
    end
end
return RI;

Example 5. The transitive reduction of the informative basis for approximate association rules extracted from the context $\mathcal{D}$ for a minimal support threshold of 2/6 and a minimal confidence threshold of 3/6 is presented in Table 3. It contains seven rules, versus ten rules in the informative basis, whereas thirty six approximate association rules are valid on the whole.

4 Experimental results

We used the four following datasets during these experiments:

- T20I6D100K[4], made up of synthetic data built according to the properties of sales data, which contains 100,000 objects with an average size of 20 items and an average size of the potential maximal frequent itemsets of six items.
- Mushrooms, that consists of 8,416 objects of an average size of 23 attributes (23 items by objects and 127 items on the whole) describing characteristics of mushrooms.
- C20D10K and C73D10K[5] which are samples of the file Public Use Microdata Samples containing data of the census of Kansas carried out in 1990. They consist of 10,000 objects corresponding to the first 10,000 listed people, each

[4] http://www.almaden.ibm.com/cs/quest/syndata.html
[5] ftp://ftp2.cc.ukans.edu/pub/ippbr/census/pums/pums90ks.zip

Generator	Closure	Closed superset	Approximate rule	Support	Confidence
{A}	{AC}	{ABCE}	A → BCE	2/6	2/3
{B}	{BE}	{BCE}	B → CE	4/6	4/5
{B}	{BE}	{ABCE}			
{C}	{C}	{AC}	C → A	3/6	3/5
{C}	{C}	{BCE}	C → BE	4/6	4/5
{C}	{C}	{ABCE}			
{E}	{BE}	{BCE}	E → BC	4/6	4/5
{E}	{BE}	{ABCE}			
{AB}	{ABCE}				
{AE}	{ABCE}				
{BC}	{BCE}	{ABCE}	BC → AE	2/6	2/4
{CE}	{BCE}	{ABCE}	CE → AB	2/6	2/4

Table 3. Transitive reduction of the informative basis for approximate association rules extracted from $\mathcal{D}$ for *minsupport* = 2/6 and *minconfidence* = 3/6.

object containing 20 attributes (20 items by objects and 386 items on the whole) for C20D10K and 73 attributes (73 items by objects and 2,178 items on the whole) for C73D10K.

Execution times (not presented here) of the generation of the bases, as for the generation of all valid association rules, are negligible compared to execution times of the frequent (closed) itemsets extraction.

Number of exact association rules extracted. The total number of valid exact association rules and the number of rules in the generic basis are presented in Table 4. No exact association rule is extracted from T10I4D100K as for this support threshold all the frequent itemsets are frequent closed itemsets: they all have different supports and are thus themselves their unique generator. Consequently, there exists no rule of the form $l_1 \Rightarrow (l_2 \setminus l_1)$ between two frequent itemsets whose closures are identical: $\gamma(l_1) = \gamma(l_2)$ that are the valid exact association rules.

Dataset	Minsupport	Exact rules	Generic basis
T10I4D100K	0.5%	0	0
Mushrooms	30%	7,476	543
C20D10K	50%	2,277	457
C73D10K	90%	52,035	1,369

Table 4. Number of exact association rules extracted.

For the three other datasets, made up of dense and correlated data, the total number of valid exact rules varies from more than 2,000 to more than 52,000, which is considerable and makes it difficult to discover interesting relationships. The generic basis represents a significant reduction of the number of extracted rules (by a factor varying from 12 to 50) and since it does not represent any loss of information, it brings a knowledge that is complete, relevant and easily usable from the point of view of the user.

Number of approximate association rules extracted. The total number of valid approximate association rules and the number of rules in the transitive reduction of the informative basis are presented in Table 5. The total number of valid approximate association rules is for the four datasets very significant since it varies of almost 20,000 rules for T20I6D100K to more than 2,000,000 rules for C73D10K. It is thus essential to reduce the set of extracted rules in order to make it usable by the user. For T20I6D100K, this basis represents a division by a factor of 5 approximately of the number of extracted approximate rules. For MUSHROOMS, C20D10K, and C73D10K, the total number of valid approximate association rules is much more important than for the synthetic data since these data are dense and correlated and thus the number of frequent itemsets is much higher. As a consequence, it is the same for the number of valid approximate rules. The proportion of frequent closed itemsets among the frequent itemsets being weak, the reduction of the informative basis for approximate rules makes it possible to reduce considerably (by a factor varying from 40 to 500) the number of extracted rules.

Dataset (Minsupport)	Minconfidence	Approximate Rules	Informative basis reduction
T10I4D100K	70%	20,419	4,004
(0.5%)	30%	22,952	4,519
MUSHROOMS	70%	37,671	1,221
(30%)	30%	71,412	1,578
C20D10K	70%	89,601	1,957
(50%)	30%	116,791	1,957
C73D10K	90%	2,053,896	5,718
(90%)	80%	2,053,936	5,718

Table 5. Number of approximate association rules extracted.

Comparing rules in the generic basis and the reduction of the informative basis to all valid rules, we checked that these bases do not contain any redundant rules. Considering the example presented in the section 1 concerning the nine approximate rules extracted the dataset MUSHROOMS, only the 6^{th} rule is generated among these nine rules in the bases. Indeed, the itemsets $\{\text{free gills}\}$ and $\{\text{free gills}, \text{eatable}, \text{partial veil}, \text{white veil}\}$ are two frequent closed itemsets of which the first is an immediate predecessor of the second and they are the only frequent closed itemsets of the interval $[\varnothing, \{\text{free gills}, \text{eatable}, \text{partial veil}, \text{white veil}\}]$. Moreover, the frequent closed itemset $\{\text{free gills}\}$ being itself its unique generator, the rule 6 belongs to the transitive reduction of the informative basis: it is the minimal non-redundant rule among these nine rules.

5 Conclusion

Using the frequent closed itemsets and their generators extracted by the algorithms Close or A-Close, we define the generic basis for exact association rules and the transitive reduction of the informative basis for approximate association rules. The union of these bases provides a non-redundant generating set for all the valid association rules, their supports and their confidences. It contains the minimal non-redundant association rules (of minimal antecedent and maximal consequent) and does not represent any loss of information: from the point of view of the user, these rules are the most useful and the most relevant association rules. All the information conveyed by the set of valid association rules is also conveyed by the union of these two bases. Two algorithms for generating the generic basis and the transitive reduction of the informative basis using the frequent closed itemsets and their generators, are also presented. These bases are also of strong interest for:

- The visualization of the extracted rules since the reduced number of rules in these bases, as well as the distinction of the exact and the approximate rules, facilitate the presentation of the rules to the user. Moreover, the absence of redundant rules in the bases and the generation of the minimal non-redundant rules are of significant interest from the point of view of the user [KMR+94].
- The identification of the minimal non-redundant association rules among the set of valid association rules extracted, using Definition 5. It is thus possible to extend an existing implementation for extracting association rules or to integrate this method in the visualization system in order to present the minimal non-redundant association rules to the user.
- The data analysis and the formal concept analysis since they do not represent any loss of information compared to the set of valid implication rules and are constituted of the most useful and relevant rules. Definition 5 of the minimal non-redundant rules being also valid within the framework of global and partial implication rules between binary sets of attributes, definitions 6 of the generic basis and 7 of the informative basis are also valid for the global and partial implication rules respectively.

Moreover, we think that this approach is complementary with approaches for selecting association rules to be vizualised, such as templates and item constraints, that help the user managing the result.
As pointed out in section 3, up to now, there does not exist any inference system with completeness and soundness properties, for inferring association rules that takes into account supports and confidences of the rules. We think that the definition of such an inference system, equivalent to the Armstrong axioms for implications, constitutes an interesting perspective of future work.

References

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. *Proc. SIGMOD conf.*, pp 207–216, May 1993.

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. *Proc. VLDB conf.*, pp 478–499, September 1994.

[Arm74] W. W. Armstrong. Dependency structures of data base relationships. *Proc. IFIP congress*, pp 580–583, August 1974.

[BAG99] R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. *Proc. ICDE conf.*, pp 188–197, March 1999.

[BG99] R. J. Bayardo, and R. Agrawal. Mining the most interesting rules. *Proc. KDD Conference*, pp 145–154, August 1999.

[BB79] C. Beeri, P. A. Bernstein. Computational problems related to the design of normal form relational schemas. Transactions on Database Systems, 4(1):30–59, 1979.

[BMS97] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlation. *Proc. SIGMOD conf.*, pp 265–276, May 1997.

[DG86] V. Duquenne and J.-L. Guigues. Famille minimale d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines*, 24(95):5–18, 1986.

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

[HF95] J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. *Proc. VLDB conf.*, pp 420–431, September 1995.

[Hec96] D. Heckerman. Bayesian networks for knowledge discovery. *Advances in Knowledge Discovery and Data Mining*, pp 273–305, 1996.

[KMR+94] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. *Proc. CIKM conf.*, pp 401–407, November 1994.

[Lux91] M. Luxenburger. Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines*, 29(113):35–55, 1991.

[Mai80] D. Maier. Minimum covers in relational database model. *Journal of the ACM*, 27(4):664–674, 1980.

[NLHP98] R. T. Ng, V. S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained association rules. *Proc. SIGMOD conf.*, pp 13–24, June 1998.

[PBTL98] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Pruning closed itemset lattices for association rules. *Proc. BDA conf.*, pp 177–196, Octobre 1998.

[PBTL99a] 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.

[PBTL99b] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. *Proc. ICDT conf.*, LNCS 1540, pp 398–416, January 1999.

[PBTL99c] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Closed set based discovery of small covers for association rules. *Proc. BDA conf.*, pp 361–381, Octobre 1999.

[PSM94] G. Piatetsky-Shapiro and C. J. Matheus. The interestingness of deviations. *AAAI KDD workshop*, pp 25–36, July 1994.

[ST96] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974, December 1996.

[SBM98] C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2(1):39–68, January 1998.

[SA95] R. Srikant and R. Agrawal. Mining generalized association rules. *Proc. VLDB conf.*, pp 407–419, September 1995.

[SVA97] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. *Proc. KDD conf.*, pp 67–73, August 1997.