$\mathcal{IGB}$: une nouvelle base générique informative des règles d'association

G. Gasmi* , S. Ben Yahia*, E. Mephu Nguifo† and Y. Slimani*

* Départment des Sciences de l'Informatique
Faculté des Sciences de Tunis
Campus Universitaire, 1060 Tunis, Tunisie.
ghada_gasmi@yahoo.fr
{sadok.benyahia,yahya.slimani}@fst.rnu.tn
† CRIL-CNRS-IUT de Lens
Rue de l'Université SP 16, 62307 Lens Cedex
mephu@cril.univ-artois.fr

Résumé

Loin des promesses affichées, les règles d'association, pouvant être extraites à partir des bases de données réelles, présentent toujours un handicap de taille en vue de leur exploitation par les utilisateurs. Ceci a suscité l'intérêt des chercheurs, qui ont essentiellement puisé dans les résultats issus de l'Analyse de Concepts Formels, pour extraire "un noyau irréductible" sous-ensemble réduit de règles, appelé base générique de règles d'association. Cependant, la plupart de ces approches souffrent d'un certain nombre d'insuffisances, tant sur le plan compacité que celui de la sémantique.

Notre contribution consiste à concilier entre la compacité de la base générique présentée à l'utilisateur et la non perte d'information. À cet effet, nous définissons une nouvelle base générique informative et un système axiomatique valide et complet permettant la dérivation de toutes les règles d'association. Par ailleurs, nous proposons une nouvelle catégorisation de règles "factuelles" et "implicatives" en vue d'améliorer la qualité d'exploitation des connaissances présentées à l'utilisateur. Les résultats des expérimentations menées sur des jeux de données réels ont montré des gains importants en terme de compacité de la base générique proposée.

Abstract

Avoiding the extraction of an overwhelming knowledge is of primary importance as it guaranties extra value knowledge usefulness and reliability. In the association rule extraction topic, lossless selection is mainly owed to the determination of an "irreducible nucleus" of more compact subset of all association rules, called generic basis, from which the remaining (redundant) association rules are generated.

A critical overview of dedicated approaches pointed out some weaknesses that the present paper tries to shed light on.

Our contribution consists in reconciling between the compactness and the information lossless of the generic basis presented to the user. For that, we define a new informative generic basis and a valid and complete axiomatic system allowing the derivation of all the association rules. In addition, we propose a new categorization of "factual" and "implicative" rules in order to improve quality of exploitation of the knowledge presented at the user. Carried out experiments on benchmark datasets pointed out important profits in term of compactness of the proposed generic basis versus those proposed in the literature.

1 INTRODUCTION

Le problème d'extraction des règles d'association a été introduit par Agrawal *et al.* [1] et concernait l'analyse des bases de transactions de vente. En effet, les règles d'association ont pour but d'identifier les articles ou les items qui sont fréquemment achetés ensemble. L'extraction des règles d'association consiste à déterminer les règles dont le *support* (*i.e.*, fréquence d'apparition) et la *confiance* (*i.e.*, force d'une règle) sont au moins égaux, respectivement, à un seuil minimal de support *minsup* et un seuil minimal de confiance *minconf*, prédéfinis par l'utilisateur. Ce problème a été décomposé ainsi [1] :

1. Extraction des itemsets fréquents ;
2. Dérivation de règles d'association à partir des itemsets fréquents.

La dernière décennie a été marquée par un effort algorithmique "obsessionnel" pour la réduction du temps de calcul de l'étape d'extraction des motifs intéressants. Le succès obtenu est essentiellement dû à des prouesses de programmation avec la conjonction de la manipulation de structures de données compactes en mémoire centrale. Ceci a été illustré par Heikki Manilla : "*La fouille de données est l'art de compter*"[3]. Cependant, il semble évident que cette frénésie a fait perdre de vue l'objectif essentiel de cette étape, *i.e.*, extraire une connaissance fiable, de taille "exploitable" pour les utilisateurs finaux. Ainsi, la quasi-totalité de ces algorithmes se sont focalisés sur l'énumération de tous les motifs – présentant une fréquence d'apparition jugée satisfaisante – maximaux ou fermés [15].

Le revers de la médaille de ce succès est que cette énumération va générer un nombre impressionnant et inexploitable de règles d'association– dont un large nombre de règles redondantes–, même pour un contexte de taille raisonnable [24, 33, 40, 29]. Par conséquent, l'utilisateur, noyé dans cette masse de connaissances, ne peut tirer profit de cette connaissance, qui ne présente aucune valeur ajoutée.

Néanmoins, ce problème a encouragé le développement d'outils de classification des règles selon leurs propriétés, de sélection de sous-ensembles

[3] "Data mining is the art of counting", lors d'un tutoriel donné à BDA'2000, Blois, France.

de règles selon des critères définis par l'utilisateur, et de visualisation de ces règles sous une forme intelligible. Cet élagage peut reposer sur des patrons définis par l'utilisateur (*user-defined templates*) [22, 23, 28], des opérateurs booléens ou *SQL-Like* [31, 32, 36]. Le nombre de règles peut être aussi réduit en les élaguant avec une information additionnelle, telle qu'une taxonomie [27, 19, 20, 35] ou une métrique d'intérêt additionnelle [5, 21] (*e.g.*, corrélation de Pearson ou le χ^2-test) ou également sur l'incorporation d'une taxonomie d'item dans un processus anthropocentrique [8]. Cependant, toutes ces approches permettent d'élaguer l'espace de règles et peuvent entraîner une perte d'information. D'autres techniques, permettant une réduction sans perte d'information, repose sur une batterie de résultats issus de la théorie de l'Analyse Formelle de Concepts (AFC) [11][4]. Le fil directeur de ces approches est de :

1. déterminer l'ensemble minimal de règles d'association (ou *base générique*) présenté à l'utilisateur tout en maximisant la quantité d'informations utiles véhiculées ;
2. disposer d'un mécanisme d'inférence qui, suite à la demande de l'utilisateur, permet de retrouver le reste de règles d'association tout en déterminant avec exactitude leurs supports et leurs confiances sans accéder à la base de données.

Un survol des différentes approches de définition de bases génériques montre que leur construction est faite d'une manière syntaxique sans accorder une importance à la sémantique des règles d'association présentées à l'utilisateur. Ainsi, la négligence du côté sémantique, peut entraîner une exploitation et une interprétation inefficaces des règles d'association. Dans ce papier, nous proposons une nouvelle base générique de règles d'association tout en tenant compte des points suivants :

- limitation des règles d'association à présenter à celles qui véhiculent le maximum d'informations utiles ;
- proposition d'un mécanisme d'inférence permettant la dérivation de toutes les règles d'association suite à la demande de l'utilisateur ;
- détermination exacte des mesures de support et de confiance des règles d'association dérivées ;
- interprétation sémantique des règles d'association présentées.

Le reste de l'article est organisé comme suit. La section 2 introduit les notions de base de l'extraction des règles d'association et leur connexion avec l'Analyse des Concepts Formels. Nous consacrons la section 3 à une revue critique de la littérature relative à l'extraction des bases génériques. À la lumière des lacunes recensées, la section 4 introduit une nouvelle base informative des règles d'association. Nous proposons aussi un mécanisme d'inférence par le biais d'un système axiomatique valide et complet, permettant de dériver toutes les règles valides pouvant être extraites à partir d'une

[4] Elle est appelée aussi Analyse des Concepts Formels.

base de données. Les résultats des expérimentations menées sur des jeux de données sont présentés dans la section 5. La conclusion et les travaux futurs font l'objet de la section 6.

2 EXTRACTION DES RÈGLES D'ASSOCIATION

Dans ce qui suit, nous allons présenter la technique d'extraction de règles d'association et nous montrons sa connexion avec les éléments fondamentaux de l'Analyse Formelle de Concepts (AFC) [9].

	A oeufs	C pain	D sucre	T beurre	W lait
1	×	×		×	×
2		×	×		×
3	×	×		×	×
4	×	×	×		×
5	×	×	×	×	×
6		×	×	×	

TAB. 1 – Un exemple de contexte d'extraction (dans la suite, nous représentons "oeufs" par A, "pain" par C, "sucre" par D, "beurre" par T et "lait" par W).

2.1 Génération des règles d'association

La formalisation du problème d'extraction de règles associatives a été introduit par Agrawal *et al.* [1]. Il est défini comme suit : Soit $\mathcal{I}$= $\{i_1, i_2, ..., i_m\}$ un ensemble de *m* littéraux appelés *items* et $\mathcal{O}$ = $\{o_1, o_2, ..., o_n\}$ une base contenant n transactions. Soit I un sous-ensemble d'items ou *itemset*, contenant *k* items, alors *I* est appelé *k-itemset*. Le support d'un itemset *I* est donné par *support(I)* = $\frac{|\{o_j \in \mathcal{O}, I \subseteq o_j\}|}{|\{o_j \in O\}|}$. Lorsque le support de *I* est supérieur ou égal un un seuil minimal *minsup*[5], fixé par l'utilisateur, *I* est dit *fréquent*.

Une *règle d'association* est définie comme une implication entre deux itemsets. Elle est de la forme suivante : $R : X \Rightarrow Y$ tel que $X \subseteq \mathcal{I}$, $Y \subseteq \mathcal{I}$ et $X \cap Y = \emptyset$. La règle *R* est dite *basée* sur l'itemset fréquent $X \cup Y$ et les deux itemsets *X* et *Y* sont appelés, respectivement, *prémisse* et *conclusion* de *R*.

[5] c'est le nombre de transactions minimal auxquelles doit appartenir un itemset *I* pour être qualifié de fréquent.

Afin de vérifier la validité d'une règle d'association R, deux mesures sont communément utilisées [1] :

- Le *support* : Le support de la règle *R*, noté *support(R)* est donné par $support(X \cup Y)$.
- La *confiance* : elle exprime la probabilité conditionnelle qu'une transaction contienne *Y* sachant qu'elle contient *X*. La confiance de la règle *R*, notée *confiance(R)*, est mesurée par le ratio $\frac{support(X \cup Y)}{support(X)}$. Une règle d'association ayant une confiance égale à 1 est appelée *règle d'association exacte*. Sinon, elle est appelée *règle d'association approximative* [33].

L'extraction des règles d'association consiste à déterminer l'ensemble des règles *valides i.e.*, dont le support et la confiance sont au moins égaux, respectivement, à un seuil minimal de support *minsup* et à un seuil minimal de confiance *minconf* prédéfinis par l'utilisateur. Une décomposition de ce problème a été présentée dans [1] :

1. Extraction des itemsets fréquents : ce sont tous les itemsets ayant un support au moins égal à *minsup*.
2. Génération des règles d'association valides basées sur l'ensemble d'itemsets fréquents préalablement extraits : ces règles sont de la forme *R : X⇒Y, X⊂ Y* et confiance(*R*) $\geq$*minconf*.

Exemple 1
Le tableau 2 montre l'ensemble de toutes les règles d'association valides extraites à partir du contexte du tableau 1 pour minsup = 0.5 et minconf = 0.6. À cet effet, nous avons utilisé l'algorithme APRIORI *[1].*
L'itemset {*sucre, beurre*} *n'est pas fréquent car son support est égal à 33*%*.*
L'itemset {*beurre, lait*} *est fréquent et permet d'engendrer les deux règles* R_{34} *:lait* $\Rightarrow$ *beurre et* R_{39} *:beurre* $\Rightarrow$ *lait (voir le tableau 2).*

2.2 Bases de l'Analyse Formelle de Concepts (AFC)

Plusieurs approches basées sur l'analyse formelle des concepts (AFC) ont été proposées pour la génération des itemsets fréquents et les règles d'association [37]. À cet effet, nous allons présenté dans ce qui suit les notions de base de l'AFC.

Contexte formel : Un contexte formel est un triplet $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, tel que $\mathcal{O}$ un ensemble fini d'objets (ou de transactions), $\mathcal{I}$ un ensemble fini d'attributs (ou items) et $\mathcal{R}$ une relation binaire (*i.e.*, $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$). Chaque couple *(o,i)* $\in \mathcal{R}$ signifie que la transaction *o* $\in \mathcal{O}$ contient l'attribut *i* $\in \mathcal{I}$. Pour le contexte illustré par le tableau 1, $\mathcal{O}$ = {1, 2, 3, 4, 5, 6} et $\mathcal{I}$ = {oeufs, pain, sucre, beurre, lait}.

Correspondance de Galois : Soit le contexte d'extraction $\mathcal{K}$ = ($\mathcal{O}$, $\mathcal{I}$, $\mathcal{R}$). Soit l'application ϕ de l'ensemble des parties de $\mathcal{O}$ (*i.e.* l'ensemble de tous

Numéro	**Règle**	**Support**	**Confiance**	**Numéro**	**Règle**	**Support**	**Confiance**
R_1	DW⇒C	0.5	1	R_2	D⇒C	0.66	1
R_3	AT⇒CW	0.5	1	R_4	TW⇒C	0.5	1
R_5	TW⇒AC	0.5	1	R_6	ACT⇒W	0.5	1
R_7	ATW⇒C	0.5	1	R_8	AT⇒C	0.5	1
R_9	TW⇒A	0.5	1	R_{10}	A⇒CW	0.66	1
R_{11}	AC⇒W	0.66	1	R_{12}	AW⇒C	0.66	1
R_{13}	A⇒C	0.66	1	R_{14}	A⇒W	0.66	1
R_{15}	T⇒C	0.66	1	R_{16}	W⇒C	0.83	1
R_{17}	D⇒CW	0.5	0.75	R_{18}	CD⇒W	0.5	0.75
R_{19}	D⇒W	0.5	0.75	R_{20}	W⇒CD	0.5	0.6
R_{21}	CW⇒D	0.5	0.6	R_{22}	W⇒D	0.5	0.6
R_{23}	C⇒D	0.66	0.66	R_{24}	C⇒W	0.83	0.83
R_{25}	C⇒A	0.66	0.66	R_{26}	C⇒AW	0.66	0.66
R_{27}	W⇒AC	0.66	0.8	R_{28}	CW⇒A	0.66	0.8
R_{29}	W⇒A	0.66	0.8	R_{30}	C⇒T	0.66	0.66
R_{31}	W⇒ACT	0.5	0.6	R_{32}	CW⇒AT	0.5	0.6
R_{33}	W⇒AT	0.5	0.6	R_{34}	W⇒T	0.5	0.6
R_{35}	W⇒AC	0.66	0.8	R_{36}	CW⇒T	0.5	0.6
R_{37}	W⇒CT	0.5	0.6	R_{38}	T⇒A	0.5	0.75
R_{39}	T⇒W	0.5	0.75	R_{40}	T⇒AC	0.5	0.75
R_{41}	T⇒ACW	0.5	0.75	R_{42}	T⇒AW	0.5	0.75
R_{43}	CT⇒A	0.5	0.75	R_{44}	CT⇒AW	0.5	0.75
R_{45}	CT⇒W	0.5	0.75	R_{46}	A⇒T	0.5	0.75
R_{47}	A⇒TW	0.5	0.75	R_{48}	A⇒CT	0.5	0.75
R_{49}	A⇒CTW	0.5	0.75	R_{50}	AC⇒T	0.5	0.75
R_{51}	AW⇒T	0.5	0.75	R_{52}	ACW⇒T	0.5	0.75

TAB. 2 – L'ensemble de toutes les règles d'association extraites du contexte d'extraction illustré par le tableau 1 (faute d'espace, nous représentons "oeufs" par A, "pain" par C, "sucre" par D, "beurre" par T et "lait" par W).

les sous-ensembles de $\mathcal{O}$), noté par *P*($\mathcal{O}$), dans l'ensemble des parties de $\mathcal{I}$, noté par *P*($\mathcal{I}$). L'application ϕ associe à un ensemble d'objets $O \subseteq \mathcal{O}$, l'ensemble des items $i \in \mathcal{I}$ communs à tous les objets $o \in O$.

$$\phi : P(\mathcal{O}) \rightarrow P(\mathcal{I})$$
$$\phi(O) = \{i \in \mathcal{I} | \forall o \in O, (o,i) \in \mathcal{R}\}$$

Soit l'application ψ de l'ensemble des parties de $\mathcal{I}$ dans l'ensemble des parties de $\mathcal{O}$. Elle associe à un ensemble d'items $I \subseteq \mathcal{I}$, l'ensemble d'objets $o \in \mathcal{O}$ communs à tous les items $i \in I$:

$$\psi : P(\mathcal{I}) \rightarrow P(\mathcal{O})$$
$$\psi(I) = \{o \in \mathcal{O} | \forall i \in I, (o,i) \in \mathcal{R}\}$$

Le couple d'applications (ϕ, ψ) définit une *correspondance de Galois* entre l'ensemble des parties de $\mathcal{O}$ et l'ensemble des parties de $\mathcal{I}$. Les applications $\omega = \phi \circ \psi$ et $\gamma = \psi \circ \phi$ sont appelées les opérateurs de fermeture de la correspondance de Galois [11].
ϕ({1, 2, 3})={pain, lait}.
ψ({oeufs, pain})={1, 3, 4, 5}.

Itemset fermé : Un itemset $I \subseteq \mathcal{I}$ est fermé si seulement si $I = \omega(I)$ [34]. *I* est un ensemble maximal d'items communs à un ensemble d'objets. Un itemset fermé *I* est fréquent si seulement si son support noté support(*I*)= $\frac{|\psi(I)|}{|\mathcal{O}|}$ $\geq$ *minsup* (*i.e.*, le seuil minimal de support).
Soit le contexte d'extraction illustré par le tableau 1. Pour minsup = 0.5, {oeufs, beurre, lait, pain} est itemset fermé fréquent de support 0.5.

Concept formel : Un concept formel est une paire $c = (O, I)$, où *O* est appelé *extension* et *I* est appelé *intension*. $\phi(O)$ = *I* et $\psi(I)$ = *O*.
La paire ({1, 3, 5}[6],{oeufs, beurre, lait, pain) est un concept formel.

Générateur minimal : Un itemset $g \subseteq \mathcal{I}$ est un générateur minimal d'un itemset fermé *I* si seulement si $\omega(g)$ = *I* et $\nexists$ *g'* $\subseteq$ *g* tel que $\omega(g')$ = *I* [4].
Soit l'itemset fermé fréquent {oeufs, beurre, lait, pain}, ses générateurs minimaux correspondants sont {oeufs, beurre} et {lait, beurre}

Itemset pseudo-fermé : Un itemset $I_1 \in \mathcal{I}$ est un itemset pseudo-fermé s'il n'est pas fermé et s'il contient les fermetures de tous ses sous-ensembles qui sont des itemsets pseudo-fermés. Ainsi, l'ensemble des itemsets pseudo-fermés $\mathcal{PF}$ = { $I_1 \in \mathcal{I} \mid I_1 \neq \omega(I_1) \wedge \forall\, I_2 \subset I_1$ tel que $I_2 \in \mathcal{PF}$, nous avons $\omega(I_2) \subset I_1$} [33].
{oeufs, pain} est un itemset pseudo fermé puisque {oeufs, pain} contient

[6] {1, 3, 5} s'écrit aussi 135

{pain} qui est la fermeture de l'itemset pseudo-fermé $\emptyset$.

Classe d'équivalence : L'ensemble d'itemsets est réparti en des classes d'équivalence contenant chacune des itemsets ayant la même fermeture. Les générateurs minimaux sont les éléments les plus petits d'une classe d'équivalence, tandis que l'itemset fermé constitue l'élément maximal de la classe [33].
Par exemple, les itemsets {oeufs}, {oeufs, pain}, {oeufs, lait} et {oeufs,pain, lait} appartiennent à une même classe d'équivalence puisqu'ils ont la même fermeture.

Treillis de Galois (Treillis de concepts) : Soit $\mathcal{C}$ l'ensemble de concepts formels extraits du contexte formel $\mathcal{K}$ = ($\mathcal{O}$, $\mathcal{I}$, $\mathcal{R}$). L'ensemble $\mathcal{C}$ doté de la relation d'ordre $\leq$ est appelé treillis de Galois, noté $\mathcal{L}_c = (\mathcal{C}, \leq)$ [3]. Une telle structure vérifie les deux propriétés suivantes :

- Un ordre partiel entre les concepts formels est défini comme suit : $\forall$ c_1, $c_2 \in$ $\mathcal{C}_\mathcal{K}$, $c_1 \leq c_2$ si seulement si intension(c_2)$\subseteq$ intension(c_1) $\vee$ extension(c_1)$\subseteq$ extension(c_2). L'ordre partiel est utilisé pour générer le graphe du treillis, représentant la relation de couverture, de la manière suivante : il existe un arc *(c_1, c_2)* si $c_1 \leq c_2$, $\nexists$ de concept c_3, tel que $c_1 \leq c_3 \leq c_2$. Le graphe est appelé diagramme de Hasse [11].
- Tous les sous-ensembles de $\mathcal{L}_c$ ont une borne inférieure appelée élément *Meet, i.e.* le plus grand minorant (*GLB*) de $\mathcal{L}_c$, et une borne supérieure appelée élément *Join*, i.e. le plus petit majorant (*LUB*) de $\mathcal{L}_c$.

Treillis de l'iceberg de Galois : Soit $\mathcal{IFF}$ l'ensemble de itemsets fermés fréquents extraits du contexte formel $\mathcal{K}$ = ($\mathcal{O}$, $\mathcal{I}$, $\mathcal{R}$). La paire $\mathcal{L}_{IFF} = (\mathcal{IFF}, \leq)$, préservant uniquement *le plus petit majorant* (*LUB*), est appelée *treillis de l'iceberg de Galois* [11]. Soit le contexte d'extraction $\mathcal{K}$ illustré par le tableau 1. La figure 1 présente le treillis de l'iceberg de Galois correspondant. Chaque noeud du treillis représente le couple (itemset fermé, support) et il est décoré par la liste de ses générateurs minimaux.

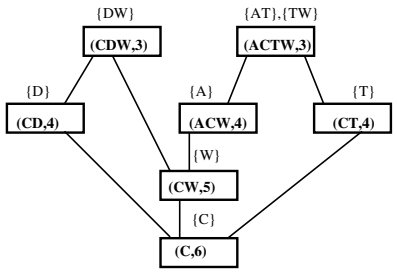

FIG. 1 – Treillis de l'iceberg de Galois associé au contexte d'extraction du tableau 1 pour *minsup* = 0.5 (Faute d'espace, nous représentons "oeufs" par A, "pain" par C, "sucre" par D, "beurre" par T et "lait" par W.

3 EXTRACTION DES RÈGLES D'ASSOCIATION NON REDONDANTES

Le problème de l'exploitation du résultat présenté à l'utilisateur est devenu primordial, étant donné le nombre élevé de règles d'association extraites à partir d'ensembles de données réels. Ceci a suscité l'intérêt des chercheurs qui ont essentiellement puisé dans les travaux issus de la théorie de l'analyse formelle des concepts (AFC), pour proposer des approches de sélection sans perte d'information des règles d'association qui véhiculent le maximum de connaissances utiles. Cette approche repose sur l'extraction d'un sous-ensemble générique, de toutes les règles d'association, appelé *base générique*, tout en satisfaisant les conditions suivantes [25] :

1. ***Dérivabilité :*** Le mécanisme d'inférence[7] (*e.g.*, un système axiomatique), permettant la dérivation des règles redondantes doit être valide (*i.e.,* le système ne permet de dériver que les règles d'association valides) et complet (*i.e.*, l'ensemble de toutes les règles valides peut être retrouvé).
2. ***Informativité :*** la base générique des règles d'association doit permettre de retrouver avec exactitude le support et la confiance des règles dérivées.

Une revue de la littérature nous a permis de classer les travaux portant sur l'extraction des bases génériques en deux catégories : avec perte d'information et sans perte d'information.

3.1 Bases génériques avec perte d'information

Les approches d'extraction de bases génériques, présentées dans la suite, se font avec perte d'information : elles ne satisfont pas l'une des deux conditions précédemment citées (*i.e.*, la dérivabilité ou l'informativité).

3.1.1 La base Guigues-Duquenne ($\mathcal{DG}$)

La base $\mathcal{DG}$ définie par Guigues et Duquenne [16] a été redéfinie par Pasquier [33] pour qu'elle soit adaptée au cadre des règles d'association. $\mathcal{DG}$ n'est pas informative [33]. La dérivabilité est assurée par les axiomes d'Armstrong ($\mathcal{AA}$) [2][8] permettant de retrouver seulement les règles d'association exactes. Néanmoins, les règles dérivées ne sont pas toujours valides. En effet, ($\mathcal{AA}$) est un mécanisme d'inférence complet mais il n'est pas valide. Par conséquent, l'approche d'extraction de la base $\mathcal{DG}$ est qualifiée d'approche avec perte d'information [26].

[7] Un rappel sur les mécanismes d'inférence est fait en annexe.

[8] Voir annexe.

3.1.2 La base propre ($\mathcal{BP}$)

Adaptant la base $\mathcal{BP}$, définie par Luxenburger [30], au cadre des règles d'association, Pasquier définit dans [33] la base propre pour les règles d'association approximatives. Afin de dériver les règles approximatives redondantes à partir de la base $\mathcal{BP}$, Pasquier propose d'appliquer le mécanisme d'inférence $\mathcal{IRF}$ (Inférence des règles basée sur la fermeture) [33][9]. Toutefois, il a été montré dans [25] que ce mécanisme d'inférence est complet mais il n'est pas valide et que la base $\mathcal{BP}$ n'est pas informative. Afin de pallier ce problème, Kryszkiewicz propose d'étendre la base $\mathcal{BP}$ par $\mathcal{DG}$ [25] et d'utiliser le couple ($\mathcal{AA}$, $\mathcal{IRF}$) pour dériver toutes les règles d'association approximatives. Néanmoins, nous signalons les limites suivantes :

- À partir du couple ($\mathcal{BP}$, $\mathcal{DG}$), nous ne pouvons déterminer que le support et la confiance des règles approximatives. Or, ça serait plus intéressant d'étendre $\mathcal{BP}$ par une base permettant la dérivation de l'ensemble de toutes les règles exactes et approximatives tout en déterminant, avec exactitude, leurs supports et leurs confiances ;
- Le couple ($\mathcal{BP}$, $\mathcal{DG}$) forme une *super-base hétérogène*. En effet, les règles de la base $\mathcal{DG}$ traduisent des implications entre itemsets pseudo-fermés et itemsets fermés fréquents. Par ailleurs, les règles de la base $\mathcal{BP}$ traduisent des implications entre itemsets fermés fréquents. Ainsi, cette hétérogénéité peut entraîner une difficulté d'interprétation des connaissances présentées à l'utilisateur.

3.1.3 Les Règles Représentatives ($\mathcal{RR}$)

Dans [24], Kryszkiewicz introduit une base générique de règles d'association, appelée *base de règles représentatives* $\mathcal{RR}$. Afin de dériver l'ensemble de toutes les règles d'association redondantes, Kryszkiewicz propose un mécanisme d'inférence appelé *opérateur de couverture* $\mathcal{C}$ [24][10]. Dans [24], il a été prouvé que la base $\mathcal{RR}$ n'est pas informative et que l'opérateur de couverture $\mathcal{C}$ est valide et complet. Par conséquent, l'extraction de $\mathcal{RR}$ est faite avec perte d'information.
Il est à noter que la base $\mathcal{RR}$ a été redéfinie par Luong [29] sous le nom de base représentative $\mathcal{BR}$, avec la différence minime suivante : la prémisse et la conclusion des règles génériques de $\mathcal{BR}$ ne sont pas nécessairement disjointes. Aussi, nous remarquons que l'axiome d'augmentation gauche et l'axiome de décomposition, proposés par Luong pour dériver les règles redondantes, forment ensemble l'opérateur de couverture $\mathcal{C}$ [14].

---

[9] Voir annexe.

[10] Voir annexe.

3.1.4 Les Règles d'association Non Redondantes ($\mathcal{RNR}$)

Dans [41], Zaki introduit une nouvelle base générique de règles d'association, appelée base de règles d'association non Redondantes $\mathcal{RNR}$. Afin de dériver l'ensemble de toutes les règles redondantes, Zaki propose d'utiliser l'axiome de transitivité de Luxenburger [30][11] ainsi que l'axiome d'augmentation d'Armstrong [2]. Toutefois, dans [12], nous avons noté que l'utilisation de ces axiomes ne garantit pas la génération de toutes les règles valides. Ainsi, le mécanisme d'inférence utilisé n'est pas complet. Cependant la base $\mathcal{RNR}$ est informative.

3.2 Bases génériques sans perte d'information

3.2.1 Base générique des règles exactes ($\mathcal{GBE}$) et Base informative des règles approximatives ($\mathcal{GBA}$)

Bastide *et al.* définissent la base générique des règles d'association exactes, notée $\mathcal{GBE}$, comme suit [4] :

Définition 1
Soient $\mathcal{IFF}$ l'ensemble d'itemsets fermés fréquents et $\mathcal{G}_c$ l'ensemble des générateurs minimaux d'un itemset fermé fréquent c. La base $\mathcal{GBE}$ est définie comme suit :
$\mathcal{GBE} = \{R : g \Rightarrow (c - g) \mid c \in \mathcal{IFF} \wedge g \in \mathcal{G}_c \wedge g \neq c\}$.

Exemple 2
Soit le contexte d'extraction du tableau 1. Pour minsup = 0.5 et minconf = 0.6, le tableau 3 est une illustration de la base $\mathcal{GBE}$.

Numéro	Règle	Support	Confiance
R_1	oeufs, beurre⇒pain, lait	0.5	1
R_2	beurre, lait⇒oeufs, pain	0.5	1
R_3	sucre, lait⇒pain	0.5	1
R_4	sucre⇒pain	0.5	1
R_5	lait⇒pain	0.83	1
R_6	beurre⇒pain	0.66	1
R_7	oeufs⇒pain, lait	0.66	1

TAB. 3 – La base $\mathcal{GBE}$ extraite du contexte d'extraction du tableau 1.

Bastide *et al.* ont également introduit une base informative des règles approximatives, notée $\mathcal{GBA}$, définie comme suit [4] :

[11] Voir annexe

Définition 2
Soient $\mathcal{IFF}$ l'ensemble d'itemsets fermés fréquents et $\mathcal{G}$ l'ensemble des générateurs minimaux. La base $\mathcal{GBA}$ est définie comme suit :
$\mathcal{GBA}$ = {R : g $\Rightarrow$ (c-g), c $\in \mathcal{IFF} \wedge$ g $\in \mathcal{G} \wedge \omega$(g) $\subset$ c $\wedge$ confiance(R) $\geq$ minconf}.

Exemple 3
Soit le contexte d'extraction du tableau 1. Pour minsup = 0.5 et minconf = 0.6, nous pouvons extraire la base $\mathcal{GBA}$ présentée par le tableau 4.

Numéro	Règle	Support	Confiance
R_8	pain⇒lait	0.83	0.83
R_9	beurre⇒oeufs, pain, lait	0.5	0.75
R_{10}	sucre⇒pain, lait	0.66	0.75
R_{11}	oeufs⇒pain, beurre, lait	0.5	0.75
R_{12}	pain⇒beurre	0.66	0.66
R_{13}	pain⇒sucre	0.66	0.66
R_{14}	pain⇒lait, oeufs	0.66	0.66
R_{15}	lait⇒pain, oeufs	0.66	0.8
R_{16}	lait⇒pain, oeufs, beurre	0.5	0.6
R_{17}	lait⇒pain, sucre	0.5	0.6

TAB. 4 – La base $\mathcal{GBA}$ extraite du contexte d'extraction illustré par le tableau 1.

Néanmoins, l'extraction de la base $\mathcal{GBA}$ souffre de la génération d'un nombre important de règles. Ce constat est renforcé par le fait que pour les bases éparses, l'extraction de $\mathcal{GBA}$ n'apporte aucun gain en terme de compacité [4]. À cet effet, Bastide *et al* ont défini la réduction transitive de la base informative des règles approximatives, notée $\mathcal{TGBA}$ en se basant sur l'axiome de transitivité de Luxenburger [30] :

Définition 3
***Réduction transitive de** $\mathcal{GBA}$*
$\mathcal{TGBA}$ = {R : g $\Rightarrow$ (c-g), c $\in \mathcal{IFF} \wedge$ g $\in \mathcal{G} \wedge \omega$(g) $\subset$ c, $\nexists$ c' $\mathcal{IFF} \wedge$ c' $\subset$ c $\wedge$ confiance(R) $\geq$ minconf}.

Exemple 4
Soit le contexte d'extraction du tableau 1. Pour minsup = 0.5 et minconf = 0.6, le tableau 5 illustre la base $\mathcal{TGBA}$.

Afin de dériver l'ensemble de toutes les règles d'association redondantes à partir du couple ($\mathcal{GBE}$, $\mathcal{GBA}$), Kryszkiewicz propose d'utiliser l'opérateur de couverture $\mathcal{C}$ [26]. Kryszkiewicz propose aussi d'appliquer l'opérateur de couverture $\mathcal{C}$ ainsi que l'axiome de transitivité de Luxenburger [30] afin de dériver les règles d'association à partir du couple ($\mathcal{GBE}$, $\mathcal{TGBA}$) [26].

Numéro	Règle	Support	Confiance
R_8	pain⇒lait	0.83	0.83
R_9	beurre⇒oeufs, pain, lait	0.5	0.75
R_{10}	sucre⇒pain, lait	0.66	0.75
R_{11}	oeufs⇒pain, beurre, lait	0.5	0.75
R_{12}	pain⇒beurre	0.66	0.66
R_{13}	pain⇒sucre	0.66	0.66

TAB. 5 – La base $\mathcal{TGBA}$ extraite du contexte d'extraction du tableau 1.

Dans [25], il a été prouvé que les couples ($\mathcal{GBE}$, $\mathcal{GBA}$) et ($\mathcal{GBE}$, $\mathcal{TGBA}$) sont extraits sans perte d'information.

3.3 Synthèse des différentes approches d'extraction de bases génériques

Dans le tableau 6, nous comparons les différentes approches d'extraction de bases génériques tout en tenant compte des points suivants :

1. la forme des règles génériques *i.e.*, déterminer la nature de la prémisse et la conclusion des règles génériques.
2. l'informativité de la base générique ;
3. la dérivabilité. Pour cela, nous étudions la validité et la complétude des mécanismes d'inférence utilisés pour dériver les règles d'association à partir d'une base générique ;
4. la nature des règles d'association dérivées par le mécanisme d'inférence.

À la lecture du tableau 6, nous constatons que :

1. les règles des bases $\mathcal{RR}$, $\mathcal{BR}$ et du couple ($\mathcal{GBE}$, $\mathcal{GBA}$) véhiculent le maximum d'informations utiles. En effet, il a été prouvé que les règles qui traduisent des implications entre les générateurs minimaux et les itemsets fermés fréquents véhiculent le maximum d'informations utiles, puisqu'elles possèdent une prémisse minimale et une conclusion maximale [33, 26] ;
2. les règles génériques du couple ($\mathcal{DG}$,$\mathcal{BP}$) sont de formes différentes. Ainsi, cette hétérogénéité peut entraîner une difficulté à interpréter les connaissances présentées à l'utilisateur ;
3. seules les bases génériques $\mathcal{RNR}$, $\mathcal{GBE}$, ($\mathcal{GBE}$,$\mathcal{GBA}$) et ($\mathcal{DG}$,$\mathcal{BP}$) sont qualifiées de bases informative ;
4. seuls les couples ($\mathcal{GBE}$,$\mathcal{GBA}$) et ($\mathcal{DG}$,$\mathcal{BP}$) sont extraits sans perte d'information. En effet, la non perte d'information est due d'une part au fait que ces deux couples sont informatifs et d'autre part, au fait que

Base	Forme des règles prém./conc.	Informativité	Dérivabilité			Règles dérivées
			Mécanisme d'inférence	Complet	Valide	
$\mathcal{DG}$	pseudo-fermé/ fermé	non	$\mathcal{AA}$	oui	non	$\mathcal{AR}$ exactes
$\mathcal{BP}$	fermé/ fermé fermé	non	$\mathcal{IRF}$	oui	non	$\mathcal{AR}$ approximatives
($\mathcal{BP}$, $\mathcal{DG}$)	(pseudo-fermé/ fermé) ;(fermé/ fermé)	oui	$\mathcal{IRF}$,$\mathcal{AA}$	oui	oui	$\mathcal{AR}$ approximatives
$\mathcal{RR}$	gen.min/ fermé	non	$\mathcal{C}$	oui	oui	$\mathcal{AR}$
$\mathcal{RNR}$	gen.min/ gen.min	oui	Augmentation +$\mathcal{TC}$	non	oui	$\mathcal{AR}$
$\mathcal{BR}^*$	gen.min/ fermé	non	augmentation décomposition**	oui	oui	$\mathcal{AR}$
$\mathcal{GBE}$	gen.min/ fermé	oui	$\mathcal{IRF}$	oui	oui	$\mathcal{AR}$ exactes
$\mathcal{GBA}$	gen.min/ fermé	non	$\mathcal{IRF}$	oui	non	$\mathcal{AR}$ approximatives
($\mathcal{GBE}$, $\mathcal{GBA}$)	gen.min/ fermé	oui	$\mathcal{C}$	oui	oui	$\mathcal{AR}$
($\mathcal{GBE}$, $\mathcal{TGBA}$)	gen.min/ fermé	oui	$\mathcal{C}$ + transitivité	oui	oui	$\mathcal{AR}$

TAB. 6 – Caractéristiques des bases génériques.
(*) : La base $\mathcal{BR}$ est équivalente à la base $\mathcal{RR}$
(**) : Le couple de ces axiomes forment ensemble l'opérateur de couverture $\mathcal{C}$

les mécanismes d'inférence utilisés pour la dérivation des règles d'association sont valides et complets ;

5. Seules les approches d'extraction des couples ($\mathcal{GBE}$,$\mathcal{GBA}$) et ($\mathcal{GBE}$,$\mathcal{TGBA}$) permettent la dérivation de toutes les règles valides exactes et approximatives sans perte d'information.

4 $\mathcal{IGB}$: UNE NOUVELLE BASE GÉNÉRIQUE EXTRAITE SANS PERTE D'INFORMATION

Le survol de la littérature a montré que seul le couple ($\mathcal{GBE}$, $\mathcal{GBA}$) est extrait sans perte d'information. Toutefois, le nombre de règles d'association générées est très important. Pour pallier cet inconvénient, Bastide *et al.* ont proposé la base $\mathcal{TGBA}$. Cependant, les lacunes sont plus profondes. En effet, la base $\mathcal{TGBA}$ ne résout pas catégoriquement le problème "quantitatif", *i.e.*, sa taille est aussi importante. De plus, la construction des bases génériques est faite d'une manière syntaxique sans accorder une importance à la sémantique des règles. Par conséquent, l'utilisateur ne dispose des bons éléments de connaissance avec la plus value requise, pour prendre pas prendre la bonne décision.

4.1 Motivations pour la proposition d'une nouvelle base générique

Commençons par rappeler la sémantique des règles d'association. Soit la règle "$R : lait, pain \Rightarrow beurre$" de support égal à 0,5 et de confiance égale à 0,8. Une interprétation sémantique pouvant être associée à ce type de connaissance est la suivante : la probabilité de trouver les items "$lait$", "$pain$" et "$beurre$" ensemble est égale à 0,5. Tandis que, la probabilité de trouver les items "$lait$", "$pain$" et "$beurre$" d'une valeur égale à 0,8 dépend de l'existence des items "$lait$" et "$pain$" ensemble.
Soit un gestionnaire qui dispose de la base de transactions illustrée par le tableau 1. Supposons que son objectif est de pousser les ventes des lots d'items, qui se vendent avec une valeur représentant au moins 50% à une valeur représentant au moins 60% des ventes totales.
Pour ce faire, il décide de faire une promotion sur un certain nombre d'items. La question qui se pose est la suivante : Quel est l'ensemble **minimal** d'items qu'il faut promouvoir ? En utilisant le couple de bases ($\mathcal{GBE}$, $\mathcal{GBA}$) de Bastide *et al.*, le gestionnaire aura les règles génériques illustrées, respectivement, par les tableaux 3 et 4.
En utilisant le couple de bases ($\mathcal{GBE}$, $\mathcal{TGBA}$) de Bastide *et al.*, le gestionnaire aura les règles génériques illustrées, respectivement, par les tableaux 3 et 5.

4.1.1 Limites de l'utilisation du couple ($\mathcal{GBE}$, $\mathcal{GBA}$)

À la lecture des règles génériques présentées par les tableaux 3 et 4, le décideur saisit par exemple que pour pousser les ventes du lot d'items {oeufs, pain, lait} d'une valeur supérieure à 50% à une valeur supérieure à 60%, il doit promouvoir :

- L'item "pain" (*c.f.*, R_{14} :pain⇒lait, oeufs de support égal à 66% et de confiance égale à 66%) ;
- L'item "lait" (*c.f.*, R_{15} :lait⇒pain, oeufs de support égal à 66% et de confiance égale à 75%) ;
- L'item "oeufs" (*c.f.*, R_7 :oeufs ⇒lait, pain de support égal à 66% et de confiance égale à 100%).

Cependant, le gestionnaire n'a pas réellement besoin de faire une promotion pour que le lot d'items {oeufs, pain, lait} se vende avec une valeur supérieure à 60%. En effet, le lot d'items {oeufs, pain, lait} se vend toujours avec une valeur égale à 66%.

D'après les tableaux 3 et 4, les ventes du lot {oeufs, pain, lait, beurre} passent d'une valeur supérieure à 50% à une valeur supérieure à 60% suite à l'une des promotions suivantes :

- L'item "oeufs" (*c.f.*, R_{11} :oeufs⇒lait, pain, beurre de support égal à 50% et de confiance égale à 75%).
- L'item "beurre" (*c.f.*, R_9 :beurre⇒lait, pain, oeufs de support égal à 50% et de confiance égale à 75%).
- L'item "lait" (*c.f.*, R_{16} :lait⇒beurre, pain, oeufs de support égal à 50% et de confiance égale à 60%).
- les items "oeufs" et "beurre" (*c.f.*, R_1 :oeufs, beurre⇒lait, pain de support égal à 50% et de confiance égale à 100%).
- les items "lait" et "beurre" (*c.f.*, R_2 :lait, beurre⇒oeufs, pain de support égal à 50% et de confiance égale à 100%).

Quelle est la connaissance additionnelle qu'apportent les règles génériques exactes, illustrées par la base $\mathcal{GBE}$, pour le décideur ? Est-ce qu'il ne peut pas se contenter des connaissances véhiculées par les règles approximatives ? En effet, les règles "oeufs, beurre⇒lait" et "beurre⇒oeufs, pain" ne répondent pas à l'objectif du gestionnaire à savoir, déterminer l'ensemble **minimal d'items** qu'il faut promouvoir pour atteindre son objectif.

4.1.2 Limites de l'utilisation du couple ($\mathcal{GBE}$, $\mathcal{TGBA}$)

À la lecture des tableaux 3et 5, le gestionnaire saisit par exemple que pour pousser les ventes du lot {oeufs, pain, lait, beurre} d'une valeur supérieure à 50% à une valeur supérieure à 60%, il peut faire l'une des promotions suivantes :

- L'item "oeufs" (*c.f.*, R_{11} :oeufs⇒lait, pain, beurre de support égal à 50% et de confiance égale à 75%).

– L'item "beurre" (*c.f.*, R_9 :beurre⇒lait, pain, oeufs de support égal à 50% et de confiance égale à 75%).
– les items "oeufs" et "beurre" (*c.f.*, R_1 :oeufs, beurre⇒lait, pain de support égal à 50% et de confiance égale à 100%).
– les items "lait" et "beurre" (*c.f.*, R_2 :lait, beurre⇒oeufs, pain de support égal à 50% et de confiance égale à 100%).

Cependant, nous remarquons que d'une part, les règles R_1 et R_2 n'ajoutent aucune connaissance additionnelle au gestionnaire. D'autre part, le fait de ne pas présenter la règle R_{16} pourrait amener le gestionnaire à prendre la mauvaise décision. En effet, imaginons que le prix de vente du lait est plus élevé que ceux du beurre et des oeufs, d'après les règles de $\mathcal{TGBA}$, le gestionnaire ne peut promouvoir que le beurre ou les oeufs alors qu'il a la possibilité d'augmenter ses recettes par la promotion du lait.
Ainsi, nous constatons que les connaissances convoyées par les couples ($\mathcal{GBE}$, $\mathcal{GBA}$) et ($\mathcal{GBE}$, $\mathcal{TGBA}$) n'apportent pas toujours de l'utilité escomptée et qu'il y a un certain nombre de règles qui n'apportent aucune information additionnelle à l'utilisateur. D'après cet exemple, nous remarquons que le gestionnaire a besoin de :

– disposer de règles d'association lui permettant de saisir que pour certains lots d'items, il n'est pas obligé de faire une promotion ;
– visualiser seulement les règles qui comportent le minimum d'items dans leurs parties prémisse.

5 $\mathcal{IGB}$: UNE NOUVELLE BASE GÉNÉRIQUE

Ainsi, les limites du couple ($\mathcal{GBE}$, $\mathcal{GBA}$) nous ont motivés à introduire une nouvelle base générique, appelée $\mathcal{IGB}$, que nous définissons comme suit [13, 14] :

Définition 4
Soient $\mathcal{IFF}$ l'ensemble des itemsets fermés fréquents et $\mathcal{G}_c$ l'ensemble des générateurs minimaux de tous les itemsets fermés fréquents inclus ou égaux à un itemset fermé fréquent I. La base $\mathcal{IGB}$ est définie comme suit :
$\mathcal{IGB}$ = {R : $g_s \Rightarrow$ (I-g_s) | $I \in \mathcal{IFF} \wedge$ I-$g_s \neq \emptyset \wedge g_s \in \mathcal{G}_c$, $c \in \mathcal{IFF} \wedge c \subseteq I \wedge$ confiance(R) $\geq$ minconf $\wedge \nexists\, g' \mid g' \subset g_s \wedge$ confiance($g' \Rightarrow$ I-g')$\geq$ minconf }.

Ainsi, les règles génériques de $\mathcal{IGB}$ représentent des implications entre des prémisses minimales et des conclusions maximales (en terme du nombre d'items). En effet, il a été prouvé dans la littérature que ce type de règles est le plus général (*i.e.*, convoyant le maximum d'information) [26, 33].

Proposition 1
Soit I un itemset fermé fréquent non vide, si support(I)$\geq$ minconf, alors on peut extraire la règle générique suivante : $\emptyset \Rightarrow$I.

Preuve 1
La proposition 1 découle de la définition 4 : c'est le cas où g_s est vide. Nous avons confiance($\emptyset \Rightarrow$I) = support(I) et support(I)$\geq$ minconf alors $\emptyset \Rightarrow$I constitue la règle générique ayant la conclusion la plus large qu'on peut construire à partir de I puisque $\nexists$ R' :X'$\Rightarrow$Y'/X' $\subset \emptyset \wedge$ I $\subset$ Y'.

À partir de la proposition 1, nous constatons que les prémisses de certaines règles génériques de la base $\mathcal{IGB}$ peuvent être vides. Ainsi, nous distinguons deux types de règles génériques [14] :

1. ***Règles factuelles*** ayant une prémisse vide. Bien que de telles règles ont été considérées dans les travaux de Kryszkiewicz [24] et de Luong [29], aucune interprétation sémantique n'a été fournie à ce type de connaissance.
2. ***Règles implicatives*** ayant une prémisse non vide.

Dans ce qui suit, nous expliquons l'intérêt d'une telle catégorisation. Reprenons le cas du gestionnaire donné dans l'exemple 1. En utilisant $\mathcal{IGB}$ comme base générique de règles d'association, le décideur obtient l'ensemble de règles illustré par le tableau 7. Ainsi, il va saisir qu'il suffit de faire une

Numéro	**Règle**	**Support**	**Confiance**
R_1	$oeufs \Rightarrow pain, lait, beurre$	0.5	0.75
R_2	$beurre \Rightarrow pain, lait, oeufs$	0.5	0.75
R_3	$sucre \Rightarrow pain, lait$	0.5	0.75
R_4	$\emptyset \Rightarrow pain, sucre$	0.66	0.66
R_5	$\emptyset \Rightarrow pain, lait, oeufs$	0.66	0.66
R_6	$\emptyset \Rightarrow pain, beurre$	0.83	0.83
R_7	$\emptyset \Rightarrow pain, lait$	0.83	0.83
R_8	$\emptyset \Rightarrow pain$	1	1

TAB. 7 – La base $\mathcal{IGB}$ extraite du contexte donné par le tableau 1.

promotion des items "$oeufs$" ou du "$beurre$" **seulement** (*c.f.*,la règle R_1 et R_2) pour que le pourcentage des ventes du lot {oeufs, pain, beurre, lait} passe de 50% à 75% (une valeur supérieure au seuil fixé 60% des ventes totales). Cependant, pour que le pourcentage des ventes du lot {oeufs, pain, lait} excède le seuil fixé, le gestionnaire n'a nul besoin de faire une promotion (*c.f.*,la règle R_5).

Ainsi, nous remarquons que les règles de la base $\mathcal{IGB}$ rejoignent l'objectif du décideur. Ceci est dû aux raisons suivantes :

– d'une part, à de la catégorisation des règles de $\mathcal{IGB}$. En effet, *les règles factuelles* traduisent des faits (*i.e.*, l'objectif matérialisé par le seuil $minconf$ est déjà atteint). En revanche, les règles implicatives traduisent des conditions qu'il faut remplir afin d'atteindre l'objectif fixé par l'utilisateur.
– d'autre part, à la considération des règles comportant les prémisses minimales seulement.

5.1 Étude de l'approche d'extraction de la base $\mathcal{IGB}$

Dans cette section, nous allons vérifier si l'approche d'extraction de $\mathcal{IGB}$ remplit les conditions de non perte d'information à savoir la dérivabilité et l'informativité (voir section 3).

5.1.1 Dérivabilité

Afin de dériver l'ensemble de toutes les règles d'association valides, nous proposons dans la suite un système axiomatique et nous prouvons qu'il est **valide** et **complet**.

a. Validité du système axiomatique Dans cette section, nous essayons de montrer que le système axiomatique proposé est valide. Ceci revient à vérifier qu'il ne permet de dériver que les règles d'association ayant des mesures de support et de confiance au moins égales aux seuils fixés par l'utilisateur.

Proposition 2
Soient $\mathcal{IGB}$ la base générique et $\mathcal{AR}$ l'ensemble de toutes les règles valides pouvant être extraites d'un contexte d'extraction $\mathcal{K}$. Le système axiomatique suivant est valide [14] :

Réflexivité conditionnelle : *Si $X \overset{s,c}{\Rightarrow} Y$[12] $\in \mathcal{IGB}$ et $X \neq \emptyset$ alors $X \overset{s,c}{\Rightarrow} Y$ est valide.*

Augmentation : *Si $R : X \overset{s,c}{\Rightarrow} Y \in \mathcal{IGB}$ et $Z \subset Y$, alors $R' : XZ \overset{s,c'}{\Rightarrow} Y - Z$ est valide ayant une confiance c' au moins égale à c.*

Décomposition : *Soit $R : X \overset{s,c}{\Rightarrow} Y$ une règle valide obtenue suite à l'application de l'un des axiomes de réflexivité conditionnelle ou d'augmentation sur l'une des règles génériques de $\mathcal{IGB}$, $Z \subset Y$ et $\omega(XZ) = XY$. La règle $R' : X \overset{s,c}{\Rightarrow} Z$ est valide.*

Preuve 2
*La preuve de validité du système axiomatique proposé repose sur le fait que toutes les règles dérivées à partir de $\mathcal{IGB}$ sont **valides** (i.e., leurs supports et confiances sont, respectivement, supérieurs à $minsup$ et $minconf$).*

A0. Réflexivité conditionnelle : *Ceci découle de la définition même de $\mathcal{IGB}$.*

A1. Augmentation : *Le support de la règle $R = s = support(XY)$ et la confiance de $R = c = \frac{support(XY)}{support(X)}$. Puisque $X \subset XZ$, alors $support(XZ) \leq support(X)$. Nous avons $Z \subset Y$, alors nous pouvons conclure que $support(R' : XZ \Rightarrow Y - Z) = support(XY) = support(R) = s$. D'autre part, nous avons $confiance(R') = \frac{support(XY)}{support(XZ)}$. Puisque $support(XZ) \leq support(X)$, alors nous déduisons que $\frac{support(XY)}{support(XZ)} \geq \frac{support(XY)}{support(X)}$. Par conséquent, $confiance(R')$ $\geq confiance(R)$ et $support(R') = support(R) = s$.*

[12] Les valeurs s et c dénotent, respectivement, les mesures de support et de confiance de la règle d'association R.

A2. Décomposition : *Nous avons montré précédemment que la règle* $R : X \Rightarrow Y - X$ *obtenue suite à une application de l'un des axiomes de réflexivité conditionnelle ou d'augmentation est valide. Étant donné les conditions suivantes :* $X \subset XZ$ *et* $\omega(XZ) = \omega(XY)$*, alors nous avons* $support(R' : X \Rightarrow Z) = support(XZ) = support(XY) = support(R)$ *et nous avons aussi* $confiance(R') = \frac{support(XZ)}{support(X)} = \frac{support(XY)}{support(X)} = confiance(R)$*. Par conséquent,* $confiance(R') = confiance(R) = c$ *et* $support(R') = support(R) = s$*.*

Remarque 1
L'axiome de réfléxivité conditionnelle a été assorti d'une contrainte supplémentaire portant sur la non nullité de la prémisse de la règle d'association. Cette contrainte a été introduite dans le cadre du respect d'une "habitude" non explicitement énoncée sur le fait que la prémisse des règles d'association est toujours non nulle.

b. Complétude du système axiomatique Dans ce qui suit, nous montrons que le système axiomatique proposé est complet. Ceci revient à montrer qu'il permet de dériver **toutes** les règles valides pouvant être extraites d'un contexte d'extraction.

Proposition 3
Le système axiomatique proposé est complet : toutes les règles pouvant être extraites du contexte d'extraction $\mathcal{K}$ *sont dérivables de* $\mathcal{IGB}$ *moyennant le système axiomatique proposé.*

Preuve 3
Soit $\mathcal{IGB}$ *la base générique de règles obtenue à partir du contexte d'extraction* $\mathcal{K}$ *pour minsup et minconf donnés.*
Soit $\mathcal{AR}$ *l'ensemble de toutes les règles valides extraites de* $\mathcal{K}$ *et* $\mathcal{IFF}$ *l'ensemble des itemsets fermés fréquents non vides.*
Soit R :X⇒Y-X $\in \mathcal{AR}$*. Nous montrons dans ce qui suit, que R peut être dérivée à partir d'une règle générique de* $\mathcal{IGB}$ *par l'application du système axiomatique proposé.*

- *Si Y* $\in \mathcal{IFF}$*, deux cas peuvent se présenter :*
  - *si* $\nexists$ *une règle R' :X'⇒Y-X'* $\in \mathcal{AR}$ *tel que X'*$\subset$*X*
    - *si support(R)< minconf alors R :X⇒Y-X* $\in \mathcal{IGB}$*. Ainsi, R :X⇒Y-X* $\in \mathcal{AR}$ *par l'application de l'axiome de réfléxivité conditionnelle*
    - *sinon il existe une règle R" :*$\emptyset$*⇒Y*$\in \mathcal{IGB}$ *telle que en appliquant l'axiome d'augmentation à R" on retrouve la règle R :X⇒Y-X.*
  - *il existe une règle R" :X"⇒Y-X"*$\in \mathcal{IGB}$ *telle que X"*$\subset$*X'* $\wedge$ *X"*$\subset$*X. En appliquant l'axiome d'augmentation à R" :X"⇒Y-X", on obtient R :X⇒Y-X.*
- *Sinon, il existerait une règle R' :X⇒Y'-X*$\in \mathcal{AR}$ *telle que Y'*$\in \mathcal{IFF}$ $\wedge$ $\omega(Y') = \omega(Y)$*. Donc il existe une règle R" :X"⇒Y'-X"*$\in \mathcal{IGB}$ *telle*

que X"⊆X. On applique alors dans un premier temps, l'axiome d'augmentation à R" :X"⇒Y'-X"(si X"⊂X) pour obtenir R' :X⇒Y'-X. Dans un deuxième temps, on applique l'axiome de décomposition à R' pour retrouver R :X⇒Y-X.

5.1.2 Informativité

Nous avons montré que les règles dérivées à partir de $\mathcal{IGB}$ moyennant le système axiomatique proposé possèdent des mesures de support et de confiance supérieures, respectivement, à $minsup$ et $minconf$. Toutefois, peut-on déterminer ces mesures avec exactitude ? Pour ce faire, nous proposons d'étudier l'informativité de la base $\mathcal{IGB}$.

Proposition 4
La base $\mathcal{IGB}$ est informative i.e., les mesures de support et de confiance des règles redondantes peuvent être déterminées avec exactitude.

Preuve 4
*L'approche d'extraction de $\mathcal{IGB}$ consiste à trouver pour **tout itemset fermé fréquent** non vide I, les plus petits générateurs minimaux g_s associés aux itemsets fermés fréquents I' inclus dans I, tels que support(I') est au moins égal à minconf. Les règles génériques de $\mathcal{IGB}$ sont alors de la forme suivante : g_s⇒(I-g_s). Ainsi, l'ensemble de tous les itemsets fermés fréquents peut être retrouvé en concaténant la prémisse et la conclusion des règles génériques de la base $\mathcal{IGB}$. Puisque, le support d'un itemset fréquent est égal à celui de l'itemset fermé fréquent le plus petit qui le contient [33], alors la fermeture[13] et le support[14] de n'importe quel itemset peut être déterminé à partir de la base $\mathcal{IGB}$.*
Par conséquent, le support et la confiance de toutes les règles d'association peuvent être déterminés avec exactitude.

5.2 Construction de base $\mathcal{IGB}$

L'algorithme GEN-IGB-FERMÉS, dont le pseudo-code est donné par l'algorithme 1, prend en entrée l'ensemble des itemsets fermés fréquents et leurs générateurs associés extraits par l'un des algorithmes permettant l'extraction des itemsets fermés[15], *e.g.*, CLOSE, A-CLOSE [33], PRINCE [18].
L'algorithme GEN-IGB-FERMÉS est basé sur la proposition 1. En effet, il vérifie pour chaque itemset fermé non vide si son support est supérieur ou égal à *minconf* (ligne 3). Si c'est le cas, alors la règle générique $\emptyset \Rightarrow I$ est ajoutée à $\mathcal{IGB}$ (ligne 7). Sinon, il parcourt tous les itemsets fermés fréquents I' inclus dans I (ligne 10). Pour ceux ayant un support au moins égal au

[13] Ceci est utilisé par l'axiome de décomposition ($\omega(XZ) = \omega(XY)$).

[14] Ceci est utilisé par l'axiome d'augmentation (déterminer le support de XZ).

[15] Pour une étude critique de ces algorithmes, prière de se référer à [39] et [37].

rapport $\frac{support(I)}{minconf}$, il parcourt sa liste de générateurs minimaux associée. Durant ce parcours, nous cherchons le plus petit générateur minimal g_s, tel qu'il n'existe pas un générateur g' déjà inséré dans la liste $L_{plus-petite-prem}$ qui soit inclus dans g_s (ligne 11). L'algorithme GEN-IGB-FERMÉS itère ensuite sur tous les éléments de la liste $L_{plus-petite-prem}$ pour créer les règles génériques de $\mathcal{IGB}$ de la forme $g_s \Rightarrow (I - g_s)$ (ligne 14).
La complexité théorique de l'algorithme GEN-IGB-FERMÉS est en $O(\mathsf{n}^2 \times \mathsf{gen} \times \mathsf{t}^2 \times \mathsf{t}^{\frac{t}{2}})$ où :

- n : nombre d'itemsets fermés fréquents ;
- t : taille moyenne d'un itemset fermé fréquent ;
- gen : nombre moyen de générateurs minimaux associés à un itemset fermé fréquent.

Données :

1. $\mathcal{IFF}$: L'ensemble des itemsets fermés fréquents I ainsi que leurs générateurs minimaux associés $\mathcal{G}_I$ et leurs supports.
2. minconf

Résultat : $\mathcal{IGB}$:Base générique informative

```
1  Début
2    Pour Chaque itemset fermé fréquent non vide I ∈ IFF Faire
3      Si (support(I) ≥ minconf) Alors
4        R= ∅ ⇒ I;
5        R.support=support(I);
6        R.confiance=support(I);
7        IGB=IGB ∪ R;
8      Sinon
9        L_plus-petite-prem={};
10       Pour Chaque I_1 considérés dans un ordre croissant de
         taille tel que support(I)/support(I_1) ≥ miconf et I_1 ⊆I Faire
11         Pour Chaque g∈ G_I1 Faire
12           Si ∄ g_s ∈ L_plus-petite-prem tel que g_s⊂ g Alors
13             L_plus-petite-prem=L_plus-petite-prem∪g;
14       Pour Chaque g_s ∈ L_plus-petite-prem Faire
15         R= g_s ⇒I-g_s;
16         R.support=support(I);
17         R.confiance=support(I)/support(g_s);
18         IGB=IGB ∪ R;
19   retourner IGB
20 Fin
```

Algorithme 1 : GEN-IGB-FERMÉS

Exemple 5
Soit le contexte d'extraction donné par le tableau 1 (page 4). Dans ce qui suit, nous illustrons le processus d'extraction de la base $\mathcal{IGB}$ à partir des itemsets fermés fréquents et leurs générateurs pour minsup =$\frac{1}{2}$ et minconf = $\frac{2}{3}$. Ainsi, l'ensemble des itemsets fermés fréquents $\mathcal{IFF}$ est comme suit : $\{ACTW, ACW, CDW, CT, CW, CD, DW, C\}$. Pour I = ACTW, nous avons support(ACTW) = $\frac{1}{2} < \frac{2}{3}$, alors $L_{plus-petit-gen} = \{T, A\}$. Ainsi, les règles $T \overset{\frac{1}{2},\frac{3}{4}}{\Rightarrow} ACW$ et $A \overset{\frac{1}{2},\frac{3}{4}}{\Rightarrow} CTW$ sont ajoutées à la base $\mathcal{IGB}$. Pour I = ACW, support(ACW) = $\frac{2}{3}$ = minconf, la règle $\emptyset \overset{\frac{2}{3},\frac{2}{3}}{\Rightarrow} ACW$ est alors ajoutée à la base $\mathcal{IGB}$. Pour I = CDW, nous avons support(CDW) = $\frac{1}{2} <$ minconf, alors $L_{plus-petit-gen} = \{D\}$. Alors la règle $D \overset{\frac{1}{2},\frac{3}{4}}{\Rightarrow} CW$ est ajoutée à la base $\mathcal{IGB}$. Pour I = CT, support(CT) = $\frac{2}{3}$ = minconf, la règle $\emptyset \overset{\frac{2}{3},\frac{2}{3}}{\Rightarrow} CT$ est alors ajoutée à la base $\mathcal{IGB}$. Pour I = CW, nous avons support(CW) = $\frac{5}{6} >$ minconf, alors la règle $\emptyset \overset{\frac{5}{6},\frac{5}{6}}{\Rightarrow} CW$ est ajoutée à la base $\mathcal{IGB}$. Pour I = CD, nous avons support(CD) = $\frac{2}{3}$ = minconf, alors la règle $\emptyset \overset{\frac{2}{3},\frac{2}{3}}{\Rightarrow} CD$ est ajoutée à la base $\mathcal{IGB}$. Pour I = C, nous avons support(C) = 1 > minconf, alors la règle $\emptyset \overset{1,1}{\Rightarrow} C$ est ajoutée à la base $\mathcal{IGB}$.

6 ÉVALUATIONS EXPÉRIMENTALES

Afin d'évaluer la compacité de la base $\mathcal{IGB}$ par rapport aux bases génériques extraites sans perte d'information (*i.e.*, ($\mathcal{GBE}$, $\mathcal{GBA}$) et ($\mathcal{GBE}$, $\mathcal{TGBA}$)), nous avons mené une série d'expérimentations sur différentes bases benchmark denses et éparses[16] utilisées fréquemment pour évaluer les performance des algorithmes de la fouille de données. Les caractéristiques de ces bases sont présentées au niveau du tableau 8.
Les différents programmes ont été implémentés en langage C sous la distribution de Linux Fedora Core 2 avec KDevelop 3. Les caractéristiques physiques de la machine utilisée sont comme suit : un PC Pentium IV ayant une fréquence d'horloge de 3.06 Ghz et une mémoire centrale de 512 Mo.

6.1 Évolution du nombre de règles de $\mathcal{IGB}$

Dans ce qui suit, nous nous intéressons à l'étude de l'évolution du nombre de règles génériques de $\mathcal{IGB}$. À cet effet, nous étudions l'évolution du nombre des deux types de règles factuelles et implicatives suite à la variation de la valeur de $minconf$. La deuxième colonne du tableau 9 illustre le nombre d'itemsets fermés fréquents ainsi que le nombre d'itemsets fréquents extraits des bases de test pour un *minsup* fixé. La quatrième, la cinquième et

[16] L'ensemble de ces bases est disponible à l'adresse suivante : *http ://fimi.cs.helsinki.fi/data*.

Base	Nature	Nombre de transactions	Nombre d'items	Nombre d'items par transaction
T10I4D100K	Eparse	100000	1000	10
RETAIL	Eparse	88162	16470	13
ACCIDENTS	Eparse	340184	572	45
MUSHROOMS	Dense	8124	128	23
CHESS	Dense	3196	76	37
CONNECT	Dense	67775	130	43
PUMSB	Dense	49046	7117	74

TAB. 8 – Caractéristiques des bases de test.

la sixième colonnes présentent respectivement, le nombre de règles implicatives, le nombre de règles factuelles et la cardinalité de la base $\mathcal{IGB}$ pour une mesure de $minconf$ illustrée par la troisième colonne. D'après les données illustrées par le tableau 9, nous pouvons remarquer que :

- Pour $minconf = minsup$, toutes les règles de $\mathcal{IGB}$ sont $factuelles$ et leur nombre est égal à celui des itemsets fermés fréquents. Ceci peut s'expliquer par le fait que les supports de tous les itemsets fermés fréquents sont supérieurs à $minconf$. Par conséquent, l'ensemble vide constitue le générateur minimal le plus petit, tel que quelque soit l'itemset fermé fréquent I, nous avons $\frac{support(I)}{support(\emptyset)} \geq minconf$. Ainsi, le nombre de règles de $\mathcal{IGB}$ est égal à celui des itemsets fermés fréquents.
- Plus la valeur de $minconf$ augmente, plus le nombre de *règles factuelles* diminue jusqu'à atteindre 0 lorsque $minconf$ =100%. Ceci est dû à la réduction de l'ensemble d'itemsets fermés fréquents ayant un support égal au moins à la valeur $minconf$. Par conséquent, le vide ne peut plus être une prémisse des règles génériques basées sur les itemsets fermés fréquents ayant un support inférieur à $minconf$.
- Exceptionnellement pour la base de test "MUSHROOMS", le nombre de règles factuelles est égal à 1 lorsque $minconf = 100\%$. En effet, pour cette base, l'itemset *"85"* apparaît dans toutes les transactions (*i.e.*, $support(85) = support(\emptyset)$). Ainsi, quelque soit la valeur de $minconf$, la règle $\emptyset \Rightarrow 85$ est toujours valide.
- En augmentant la valeur de $minconf$, le nombre de *règles implicatives* n'évolue pas d'une manière croissante ou décroissante. Nous remarquons que pour certaines variations de $minconf$, le nombre des *règles implicatives* augmente et pour d'autres variations ce nombre diminue. En effet, le nombre de règles implicatives est déterminé par le nombre d'itemsets fermés fréquents ayant un support inférieur à $minconf$ et la cardinalité de leurs listes $L_{plus-petite-prem}$ respectives (*c.f.*, Algorithme 1, page 22). Cependant, cette cardinalité est déterminée par la taille de l'itemset fermé fréquent.
- Pour $minconf$ = 100%, le nombre de règles génériques de $\mathcal{IGB}$ est égal

Base ($minsup\%$)	|IFF| vs |IF|	$minconf$ %	|règles implica-| tives (RI)	|règles factu-| elles (RF)	|RI|+|RF|
T10I4D100K	1074 vs	00,50	0	1074	1074
(00,50)	1074	01,00	1210	385	1595
		10,00	1188	0	1188
		50,00	606	0	606
		100,00	0	0	0
RETAIL	581 vs	00,50	0	580	580
(00,50)	581	01,00	598	159	757
		10,00	418	9	427
		50,00	350	3	353
		100,00	0	0	0
ACCIDENTS	530 vs	70,00	0	530	530
(70,00)	530	80,00	715	149	864
		90,00	557	31	588
		100,00	0	0	0
MUSHROOMS	427 vs	30,00	0	427	427
(30,00)	2736	50,00	922	45	967
		70,00	954	12	966
		90,00	794	5	799
		100,00	557	1	558
CHESS	1194 vs	87,00	0	1194	1194
(87,00)	1575	89,00	1734	689	2423
		91,00	2293	362	2655
		93,00	2573	193	2766
		95,00	2681	73	2754
		97,00	1257	29	1286
		100,00	342	0	342
CONNECT	810 vs	95,00	0	809	809
(95,00)	2198	96,00	2092	480	2572
		97,00	2154	284	2438
		98,00	2463	135	2598
		99,00	2256	28	2284
		100,00	682	0	682
PUMSB	1466 vs	90,00	0	1466	1466
(90,00)	2609	92,00	3480	216	3696
		94,00	5232	216	5448
		96,00	5154	54	5208
		98,00	3330	10	3340
		100,00	1118	0	1118

TAB. 9 – Évolution du nombre de règles de $\mathcal{IGB}$ *vs* la variation de $minconf$.

à 0 lorsque nous considérons les bases de test éparses. Ceci peut être justifié par le fait pour cette mesure de $minconf$, la liste des plus petits générateurs minimaux associée à un itemset fermé fréquent I est constituée des générateurs minimaux de la classe d'équivalence de I. Puisque, nous considérons des bases de test éparses, alors les générateurs minimaux et les itemsets fermés fréquents d'une classe d'équivalence sont confondus.

– Lorsque $minsup \leq minconf$, le nombre de règles génériques de $\mathcal{IGB}$ est égal à celui des itemsets fermés fréquents puisque dans ce cas, les supports de tous les itemsets fermés fréquents sont au moins égaux à $minconf$. Ainsi, toutes les règles générées sont factuelles.

6.2 Évolution du nombre de règles des bases génériques

Dans cette section, nous comparons le nombre de règles de la base $\mathcal{IGB}$ à celui des règles des différentes bases génériques extraites sans perte d'information (*i.e.*, ($\mathcal{GBE}$, $\mathcal{GBA}$) et ($\mathcal{GBE}$, $\mathcal{TGBA}$)). Nous comparons aussi les taux de compacité des différents bases génériques par rapport au nombre de toutes les règles valides $\mathcal{AR}$ pouvant être extraites par l'utilisation de l'algorithme APRIORI [1]. Ce taux est mesuré comme suit :

$$\mathsf{comp}_{base} = 1 - (\frac{tailleBase}{taille\mathcal{AR}}).$$

À la lecture de la figure 2 et des tableaux 10 et 11, nous constatons que :

– Pour les bases éparses, l'extraction du couple ($\mathcal{GBE}$, $\mathcal{GBA}$) n'apporte aucun gain en terme de compacité. Ceci peut être justifié par le fait que l'ensemble d'itemsets fermés fréquents est confondu avec celui de générateurs minimaux et celui d'itemsets fréquents. Ainsi, construire des implications entre les générateurs minimaux et les itemsets fermés fréquents revient à construire des implications entre les itemsets fréquents.

– Pour les bases de test éparses, lorsque $minconf = 100\%$, le nombre de règles de toutes les bases génériques est égal à 0. En effet, pour cette valeur de $minconf$, $\mathcal{IGB}$, ($\mathcal{GBE}$, $\mathcal{GBA}$) et ($\mathcal{GBE}$, $\mathcal{TGBA}$) sont constitués de règles représentant des implications entre les générateurs minimaux et l'itemset fermé fréquent d'une même classe d'équivalence. Étant donné que l'ensemble d'itemsets fermés fréquents est confondu avec celui de générateurs minimaux et que la prémisse et la conclusion d'une règle doivent être disjoints, alors le nombre de règles générées est égal 0.

– La réduction transitive $\mathcal{TGBA}$ a réduit le nombre de règles génériques approximatives de $\mathcal{GBA}$. En effet, les règles génériques de $\mathcal{TGBA}$ traduisent des corrélations entre un générateur minimal g et les itemsets fermés fréquents successeurs immédiats de la fermeture de g. Cependant, les règles génériques de la base $\mathcal{GBA}$ traduisent des corrélations

entre un générateur minimal g et les itemsets fermés fréquents successeurs de la fermeture de g.
– La réduction transitive de la base informative $\mathcal{TGBA}$ a permis de réduire le nombre de règles génériques. Toutefois, la base $\mathcal{IGB}$ présente un taux de compacité qui est meilleur que celui de $\mathcal{TGBA}$.

6.3 Temps d'extraction des bases génériques sans perte d'information

Dans le tableau 12, nous présentons dans les colonnes 3 et 4, respectivement, le temps d'extraction de la base $\mathcal{IGB}$ et du couple ($\mathcal{GBE}$, $\mathcal{TGBA}$) à partir des bases de test données par le tableau 8. À cet effet, nous avons utilisé l'algorithme PRINCE [18]. À la lecture de ce tableau, nous pouvons remarquer que :

– dans les cas pratiques, le temps d'extraction des bases génériques à partir des collections de test utilisées est raisonnable malgré la taille importante de certaines d'entre elles ; Ce qui va en faveur de l'utilisation de $\mathcal{IGB}$ ou ($\mathcal{GBE}$, $\mathcal{TGBA}$) sur des bases de taille plus grande que celle des bases utilisées ;
– les temps d'extraction de $\mathcal{IGB}$ et ($\mathcal{GBE}$, $\mathcal{TGBA}$) sont presque identiques, avec un léger avantage pour $\mathcal{IGB}$. Cet avantage s'explique par le nombre de règles génériques qui est plus petit pour $\mathcal{IGB}$;
– le temps d'extraction est fonction de la taille de la collection de test et des mesures de support et de confiance. Lorsque la confiance augmente, le temps d'extraction de ($\mathcal{GBE}$, $\mathcal{TGBA}$) diminue, du fait de la diminution du nombre de règles génériques valides. Tandis qu'en augmentant la confiance, le temps d'extraction de la base $\mathcal{IGB}$ ne décroît pas nécessairement, mais reste inférieur à celui de ($\mathcal{GBE}$, $\mathcal{TGBA}$). En effet, les temps d'extraction de la base $\mathcal{IGB}$ évoluent proportionnellement au nombre de règles de $\mathcal{IGB}$, plus particulièrement par rapport au nombre de règles implicatives.

7 CONCLUSION

Le développement d'algorithmes d'extraction d'itemsets fermés fréquents de plus en plus efficaces a amélioré l'exploration de bases de données très volumineuses. Par conséquent, le nombre de connaissances qui seront présentées à l'utilisateur est prohibitif. Ceci a motivé des travaux pour la proposition d'approches d'extraction de sous-ensembles réduits de règles appelés *bases génériques* contenant les règles qui véhiculent le maximum d'information utiles. Une étude bibliographique nous a permis de constater que seules les approches d'extraction de ($\mathcal{GBE}$, $\mathcal{GBA}$) et de ($\mathcal{GBE}$, $\mathcal{TGBA}$) , proposées par Bastide *et al.*, peuvent être qualifiées d'approches sans perte d'information. Toutefois, nous avons montré que ces couples de bases ne peuvent pas

Base $minsup\%$	$minconf\%$	$\mathcal{IGB}$	$(\mathcal{GBE}, \mathcal{TGBA})$	$(\mathcal{GBE}, \mathcal{GBA})$	$\mathcal{AR}$
T10I4D100K	00,50	1074	1800	2216	2216
00,50	01,00	1595	1606	2216	2216
	10,00	1188	1195	2172	2172
	50,00	606	607	1145	1145
	100,00	0	0	0	0
RETAIL	00,50	580	1085	1382	1382
00,50	01,00	757	911	1334	1334
	10,00	427	560	770	770
	50,00	353	405	438	438
	100,00	0	0	0	0
ACCIDENTS	70,00	529	1927	8226	8226
70,00	80,00	864	1833	6445	6445
	90,00	588	1323	3407	3407
	100,00	0	0	0	0
MUSHROOMS	30,00	427	2484	7623	94894
30,00	50,00	967	2087	5761	79437
	70,00	966	1859	2159	58010
	90,00	799	1310	2159	24408
	100,00	558	557	557	8450
CHESS	87,00	1194	5278	31538	42740
87,00	89,00	2423	5278	29704	40451
	91,00	2655	5201	26147	36098
	93,00	2766	5134	21350	29866
	95,00	2754	4654	14373	20312
	97,00	1286	3228	7695	10830
	100,00	342	342	342	418
CONNECT	95,00	809	3734	25336	77816
95,00	96,00	2140	3734	18470	73869
	97,00	2438	3731	18470	60101
	98,00	2598	3484	11717	41138
	99,00	2284	3153	5250	19967
	100,00	682	682	682	2260
PUMSB	90,00	1466	8878	36640	71488
90,00	92,00	3696	8872	35142	68465
	94,00	5448	8824	29028	56349
	96,00	5208	7982	18350	35013
	98,00	3340	5720	8014	14692
	100,00	1118	1118	1118	1190

TAB. 10 – Variation du nombre de règles d'association *vs* la variation de $minconf$.

Base $minsup\%$	$minconf$ %	comp$_{\mathcal{IGB}}$ %	comp$_{(\mathcal{GBE},\mathcal{TGBA})}$ %	comp$_{(\mathcal{GBE},\mathcal{GBA})}$ %
T10I4-	00,50	51,54	18,77	00,00
D100K	01,00	28,03	27,52	00,00
00,50	10,00	45,41	44,98	00,00
	50,00	47,08	53,01	00,00
	100,00	00,00	00,00	00,00
RETAIL	00,50	58,04	21,49	00,00
00,50	01,00	43,26	31,70	00,00
	10,00	44,55	27,27	00,00
	50,00	19,41	7,53	00,00
	100,00	00,00	00,00	00,00
ACCI-	70,00	93,57	76,57	00,00
DENTS	80,00	86,60	71,55	00,00
70,00	90,00	82,75	61,16	00,00
	100,00	00,00	00,00	00,00
MUSH-	30,00	99,56	97,38	91,97
ROOMS	50,00	98,79	96,37	92,75
30,00	70,00	98,34	92,38	92,21
	100,00	93,40	93,40	93,40
CHESS	87,00	97,21	87,65	26,20
87,00	89,00	94,02	86,95	27,57
	91,00	92,71	85,59	27,57
	93,00	90,74	82,80	28,52
	95,00	86,45	77,08	29,24
	97,00	88,13	70,19	28,94
	100,00	18,19	18,19	18,19
CONNECT	95,00	98,97	95,20	67,45
95,00	96,00	97,11	94,94	67,81
	97,00	95,95	93,79	69,27
	98,00	93,69	91,53	71,52
	99,00	88,57	84,20	26,28
	100,00	69,83	69,83	69,83
PUMSB	90,00	97,95	87,58	48,74
90,00	92,00	94,61	87,04	48,68
	94,00	90,34	84,34	48,49
	96,00	85,13	77,20	47,60
	98,00	77,34	61,06	45,46
	100	6,06	6,06	6,06

TAB. 11 – Comparatif des taux de compacité des bases génériques.

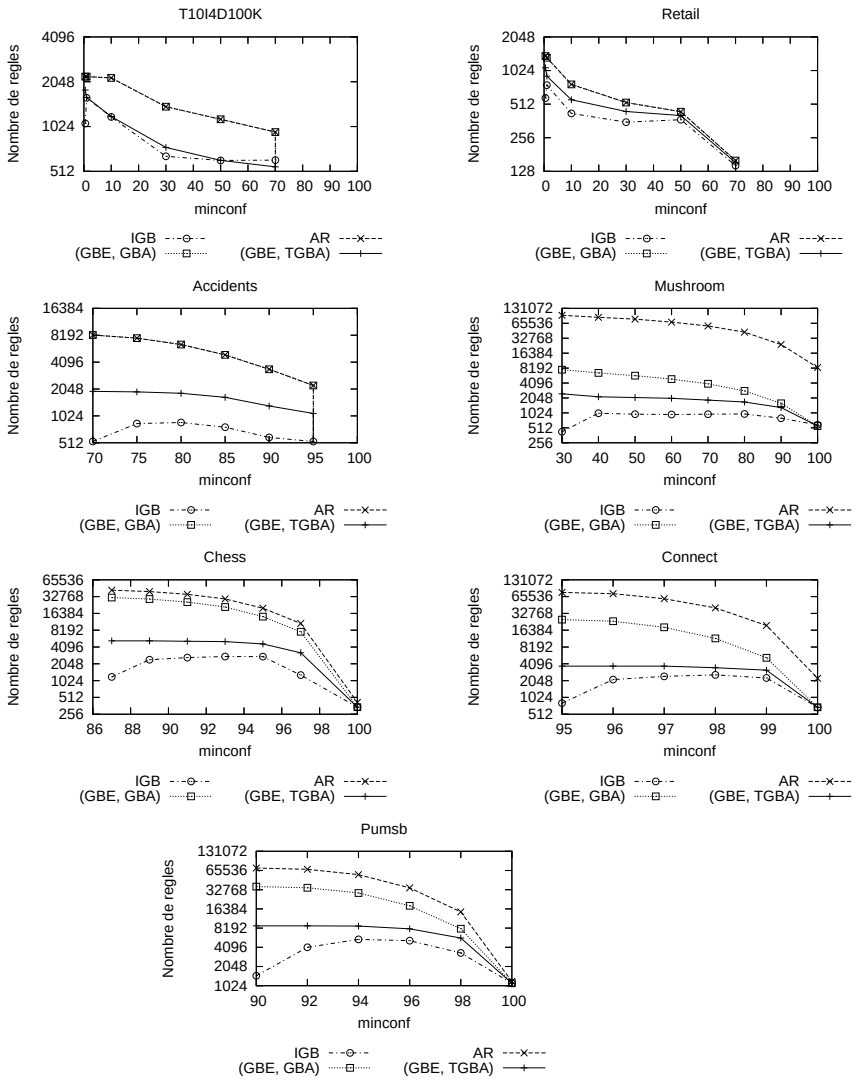

FIG. 2 – Évolution du nombre de règles d'association vs. variation de $minconf$.

Base $minsup\%$	$min-$ conf	$T_{\mathcal{IGB}}$ sec.	$T_{(\mathcal{GBE},\mathcal{TGBA})}$ sec.	$\frac{T_{\mathcal{IGB}}}{T_{(\mathcal{GBE},\mathcal{TGBA})}}$
T10I4-	00,50	3,112	3,500	0,88
D100K	01,00	3,200	3,489	0,91
00,50	10,00	3,301	3,370	0,97
	50,00	3,100	3,310	0,93
	100,00	2,925	3,260	0,89
RETAIL	00,50	10,651	11,100	0,95
00,50	01,00	10,664	11,010	0,96
	10,00	10,751	10,991	0,97
	50,00	10,600	10,800	0,98
	100,00	10,529	10,734	0,98
ACCI-	70,00	17,425	19,200	0,90
DENTS	80,00	17,511	19,280	0,91
70,00	90,00	17,530	18,900	0.92
	100,00	17,400	18,700	0.93
MUSH-	30,00	0,440	0,480	0,91
ROOMS	50,00	0,451	0,452	0,99
30,00	70,00	0,452	0,449	1,00
	100,00	0,430	0,437	0,98
CHESS	87,00	0,492	0,591	0,83
87,00	89,00	0,498	0,582	0,85
	91,00	0,502	0,552	0,90
	93,00	0,500	0,549	0,91
	95,00	0,498	0,529	0,94
	97,00	0,496	0,527	0,94
	100,00	0,495	0,499	0,99
CONNECT	95,00	4,280	4,322	0,99
95,00	96,00	4,287	4,317	0,99
	97,00	4,293	4,305	0,99
	98,00	4,275	4,301	0,99
	99,00	4,220	4,297	0,98
	100,00	4,200	4,250	0,98
PUMSB	90,00	7,120	7,171	0,99
90,00	92,00	7,139	7,168	0,99
	94,00	7,154	7,159	0,99
	96,00	7,132	7,134	0,99
	98,00	7,123	7,128	0,99
	100	7,105	7,100	1,00

TAB. 12 – Temps d'extraction des bases génériques.

constituer un outil d'aide à la décision présentant une plus-value cognitive *i.e.*, ils contiennent des règles d'association qui n'ajoutent aucune connaissance additionnelle pour l'utilisateur. À cet effet, nous avons proposé une nouvelle base générique informative $\mathcal{IGB}$ permettant de déterminer exactement le support et la confiance des règles dérivées via un système axiomatique valide et complet. Contrairement à la catégorisation classique des règles d'association (*i.e.*, exactes et approximatives), nous avons proposé deux nouveaux types de règles (i.e., factuelles et implicatives) et nous avons montré leur intérêt dans le processus de prise de décision. Les expérimentations réalisées sur des bases benchmarks ont montré que la base $\mathcal{IGB}$ apporte d'importants gains en terme de compacité par rapport aux bases génériques extraites sans perte d'information.
Les perspectives de ce travail concernent :

- L'intégration de la base $\mathcal{IGB}$ dans un environnement de visualisation [38]. Ceci permet d'améliorer l'interaction Homme/Machine où l'utilisateur aura des informations additionnelles relatives à une règle particulière (*e.g.,*, règles redondantes à la règle générique sélectionnée) ;
- La définition de nouvelles bases génériques adéquates aux "représentations succinctes". En effet, un courant émergeant dans l'algorithmique d'extraction des règles d'association préconise l'extraction efficace de ce type de représentations, e.g., le système succinct des générateurs minimaux [10, 17], les motifs essentiels[7] et les motifs non dérivables [6]. A cet effet il serait intéressant de préconiser la définition d'un système axiomatique générique, qui pourrait être instancié pour chaque base générique extraite à partir d'une représentation succincte ;
- la prise en considération des termes négatifs. En effet, les règles d'association ont été introduites pour exprimer des corrélations entre les occurrences des items. Toutefois, l'utilisateur a besoin aussi d'une technique qui exprime la corrélation entre la non occurrence des items. Le nombre de règles d'association présentées à l'utilisateur sera évidement très important nécessitant ainsi une sélection d'un sous-ensemble réduit de règles. Pour cela, nous proposons d'étudier la possibilité d'application des approches d'extraction des bases génériques pour les règles d'association positives et négatives.
- L'étude du comportement des bases génériques suite à la variation de $minsup$ et de $minconf$ peut être intéressante. En effet, ceci peut guider l'utilisateur à choisir la meilleure combinaison ($minsup$, $minconf$) pour laquelle le nombre de règles génériques de $\mathcal{IGB}$ soit le plus compacte.

REMERCIEMENTS

Les auteurs tiennent à remercier les relecteurs anonymes pour leurs remarques constructives. Ce travail est partiellement financé par le projet Franco-

Tunisien CMCU 05G1412.

RÉFÉRENCES

[1] R. Agrawal, T. Imielinski et A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Intl. Conference on Management of Data, Washington, USA*, pages 207–216, June 1993.

[2] W.W. Armstrong. Dependency structures of database relationships. In *IFIP Congress*, pages 580–583, September 1974.

[3] M. Barbut et B. Monjardet. *Ordre et classification. Algèbre et Combinatoire*. Hachette, Tome II, 1970.

[4] Y. Bastide, N. Pasquier, R. Taouil, L. Lakhal et G. Stumme. Mining minimal non-redundant association rules using frequent closed itemsets. In *Proceedings of the Intl. Conference DOOD'2000, LNCS, Springer-Verlag*, pages 972–986, July 2000.

[5] S. Brin, R. Motawni et C. Silverstein. Beyond market baskets : Generalizing association rules to correlation. In *Proceedings of the SIGMOD, Tucson, Arizona (USA)*, pages 265–276, May 1997.

[6] T. Calders et B. Goethals. Mining all non-derivable frequent itemsets. In T. Elomaa, H. Mannila et H. Toivonen, éditeurs, *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, PKDD 2002, LNCS, volume 2431, Springer-Verlag, Helsinki, Finland*, pages 74–85, August 2002.

[7] A. Casali, R. Cicchetti, L.Lakhal et S. Lopes. Motifs essentiels et inférence des fréquences. In *Proceedings 20ème Journées Bases de Données Avancées BDA'2004, Montpellier (France)*, pages 535–554, November 2004.

[8] O. Couturier, E. Mephu Nguifo et B. Noiret. Recherche de règles d'associations hiérarchiques par une approche anthropocentrée. *Revue des Nouvelles Technologies de l'Information (special issue of Intl. Conference of Journées francophones d'Extraction et Gestion des Connaissances(EGC'2004), Clermont-Ferrand,France, Cepaduès-Editions*, (1-2) :567–572, 19-23 January 2004.

[9] B.A Davey et H.A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2002.

[10] G. Dong, C. Jiang, J. Pei, J. Li et L. Wong. Mining succinct systems of minimal generators of formal concepts. In *Proceedings of 10th International Conference on Database Systems for Advanced Applications, Beijing, China*, pages 175–187, April 2005.

[11] B. Ganter et R. Wille. *Formal Concept Analysis*. Springer-Verlag, 1999.

[12] G. Gasmi. Proposition d'une nouvelle base générique de règles d'association. Rapport de mastère, Faculté des sciences de Tunis, Faculté des sciences de Tunis, Département Informatique, July 2005.

[13] G. Gasmi, S. BenYahia, E. M. Nguifo et Y. Slimani. Discovering "factual" and "implicative" generic association rules. In *Proceedings of the Intl. Conference CAP'05, Nice, France. Presses Universitaires de Grenoble*, pages 329–344, Juin 2005.

[14] G. Gasmi, S. BenYahia, E. M. Nguifo et Y. Slimani. A new informative generic base of association rules. In *Proceedings of the Ninth Pacific Asia Knowledge Discovery and Data Mining Conference (PAKDD'05), Hanoi, Vietnam*, pages 81–90, May 2005.

[15] B. Goethals et M. J. Zaki. FIMI'03 : Workshop on frequent itemset mining implementations. In B. Goethals et M. J. Zaki, éditeurs, *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03)*, volume 90 of *CEUR Workshop Proceedings, Melbourne, Florida, USA*, 19 November 2003.

[16] J.L. Guigues et V. Duquenne. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines*, (95) :5–18, 1986.

[17] T. Hamrouni, S. Ben Yahia et E. Mephu Nguifo. PRINCE-IMG : towards efficient extraction of non-redundant generic association rule bases. In *submitted to the ACM-SIGART PODS International Conference, Chiacago, USA*, June 2005.

[18] T. Hamrouni, S. Ben Yahia et Y. Slimani. PRINCE : Extraction optimisée des bases génériques de règles sans calcul de fermetures. In *Proceedings of the Intl. INFORSID Conference, Hermès, (to appear), Grenoble, France*, pages 353–368, May 2005.

[19] J. Han et Y. Fu. Discovery of multiple-level association rules from large databases. In *Proceedings of the VLDB Conference*, pages 420–431, 1995.

[20] J. Hipp, A. Myka, R. Wirth et U. Guentzer. A new algorithm for faster mining of generalized association rules. In *LNAI 1510, Springer-verlag*, 1998.

[21] E. Huellermeir. Implication-based fuzzy association rules. In *Proceedings of the PKDD'2001 : Principles of Data Mining and Knowledge Discovery, LNAI 2168, Springer-verlag, Freiburg, Germany*, pages 241–252, September 2001.

[22] D. Keim et H. Kriegel. Visualization techniques for mining large databases : A comparison. *IEEE Transactions on Knowledge and Data Engineering*, 6(8) :923–938, December 1996.

[23] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen et A. I. Verkamo. Finding interesting rules from large sets of discovered associa-

tion rules. In *Proceedings of the 3rd International Conference on Information and Knowledge Management (CIKM'94),ACM Press*, pages 401–407, November 1994.

[24] M. Kryszkiewicz. Representative association rules. In *Research and Development in Knowledge Discovery and Data Mining. Proc. of Second Pacific-Asia Conference (PAKDD). Melbourne, Australia*, 1998.

[25] M. Kryszkiewicz. Concise representations of association rules. In D. J. Hand, N.M. Adams et R.J. Bolton, éditeurs, *Proceedings of Pattern Detection and Discovery, ESF Exploratory Workshop, London, UK*, volume 2447 of *Lecture Notes in Computer Science*, pages 92–109. Springer-Verlag, September 2002.

[26] M. Kryszkiewicz. *Concise Representations of frequent patterns and association rules*. PhD thesis, Institute of Computer Science Warsaw University of Technology, 2002.

[27] Ch. Latiri et S. BenYahia. Textmining : Generating association rules from textual data. In *Proceedings of the XIXéme Congrés Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID'2001), Geneve, Switzerland*, pages 27–39, 29–30 May 2001.

[28] B. Liu, W. Hsu, K. W ang et S. Che. Visually aided exploration of interesting association rules. In *Proceedings of the 3rd Intl. Conf. on Research and Development in Knowledge Discovery and Data mining (PAKDD'99), LNCS,Vol. 1574, Springer-verlag*, pages 380–389, April 1999.

[29] V. Phan Luong. Raisonnement sur les règles d'association. In *Proceedings 17ème Journées Bases de Données Avancées BDA'2001, Agadir (Maroc), Cépaduès Edition*, pages 299–310, November 2001.

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

[31] R. Meo, G. Psaila et S. Ceri. A new sql-like operator for mining association rules. In *Proceedings of the VLDB Conference*, pages 122–133, 1996.

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

[33] N. Pasquier. *Data Mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données*. Doctorat d'université, Université de Clermont-Ferrand II, France, 2000.

[34] N. Pasquier, Y. Bastide, R. Touil et L. Lakhal. Pruning closed itemset lattices for association rules. In M. Bouzeghoub, éditeur, *Proceedings of 14th Intl. Conference Bases de Données Avancées, Hammamet, Tunisia*, pages 177–196, 26–30 October 1998.

[35] R. Skirant et R. Agrawal. Mining generalized association rules. Research report, Almaden Research Center, IBM Research Division, San Jose, USA, 1995.

[36] R. Srikant, Q. Vu et R. Agrawal. Mining association rules with item constraints. In *Proc. of the 3rd Intl. Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California*, pages 67–73, August 1997.

[37] S. Ben Yahia, T. Hamrouni et E. M. Nguifo. Frequent closed itesmet based algorithms : A thorough structural and analytical survey. *SIGKDD Explorations. Newsletter of the ACM Special Interset Group on Knowledge Discovery and Data Mining*, 8 :93–104, June 2006.

[38] S. Ben Yahia et E. M. Nguifo. Emulating a cooperative behavior in a generic association rule visualization tool. In *Proceedings of 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04), Boca Raton, Florida*, pages 148–155, November 2004.

[39] S. Ben Yahia et E. Mephu Nguifo. Approches d'extraction de règles d'association basées sur la correspondance de galois. *ISI (Ingénierie des Systèmes d'Information), Hermès-Lavoisier*, 34(9) :23–55, 2004.

[40] M. J. Zaki. Generating non-redundant association rules. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,Boston, MA*, pages 34–43, August 2000.

[41] M. J. Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery : An International Journal(DMKDJ'04)*. pages 223–248, November 2004.

ANNEXE
MÉCANISMES D'INFÉRENCE

Dans ce qui suit, nous présentons les différents mécanismes d'inférence de règles associatives proposés dans la littérature. Ces mécanismes se définissent comme étant les méthodes de dérivation (ou d'inférence) de règles à partir d'autres règles [26].

I)Les axiomes d'Armstrong ($\mathcal{AA}$)

Le système axiomatique d'Armstrong est utilisé uniquement dans le cadre des règles exactes [2] et il est constitué des axiomes suivants :

1. ***Augmentation*** : Si confiance$(R : X \Rightarrow Y) = 1$ alors, confiance($R' : XZ \Rightarrow Y) = 1$.
2. ***Transitivité*** : Si confiance$(R : X \Rightarrow Y) = 1$ et confiance$(R' : Y \Rightarrow Z)$ =1 alors, confiance$(R" : X \Rightarrow Z) = 1$.

II)La transitivité de la confiance ($\mathcal{TC}$)

Ce mécanisme, introduit par Luxenburger [30], permet de déterminer la confiance d'une règle à partir de celles d'autres règles. Contrairement à l'axiome de transitivité d'Armstrong, $\mathcal{TC}$ est utilisé aussi bien pour les règles exactes que pour les règles approximatives.

Soient X, Y et Z trois itemsets fréquents tels que $X \subset Y \subset Z$. Le support et la confiance de la règle $R" : X \Rightarrow (Z - X)$ peuvent être déterminés à partir des mesures de validité des règles $R : X \Rightarrow (Y - X)$ et $R' : Y \Rightarrow (Z - Y)$ comme suit :

1. support($R" : X \Rightarrow Z - X$) = support($R' : Y \Rightarrow Z - Y$).
2. confiance($R" : X \Rightarrow Z - X$) = confiance($R : X \Rightarrow Y - X$)$\times$ confiance($R' : Y \Rightarrow Z - Y$).

Ainsi, la confiance d'une règle dérivée est toujours inférieure ou égale aux confiances des règles à partir desquelles elle a été inférée.

III)Opérateur de couverture ($\mathcal{C}$)

Dans [24], Kryszkiewicz introduit un opérateur syntaxique de dérivation de règles associatives défini comme suit :

Définition 5
Soient X, Y, Z et V quatre itemsets fréquents et $R : X \Rightarrow Y$ une règle valide. L'ensemble de règles dérivé suite à l'application de l'opérateur de couverture sur une règle R, noté la règle $\mathcal{C}(R)$ est défini comme suit : $\mathcal{C}(R : X \Rightarrow Y)$= $\{R' : X \cup Z \Rightarrow V \mid Z \subseteq Y \wedge V \subseteq Y \wedge (Z \cap V) = \emptyset \wedge V \neq \emptyset\}$.

À partir d'une règle $R : X \Rightarrow Y$, on peut dériver, grâce à l'opérateur de couverture $\mathcal{C}$, $3^m - 2^m$ règles, où m désigne la cardinalité de l'itemset fréquent Y. Soit $\mathcal{R}$ l'ensemble de toutes les règles qui couvrent une règle associative R' (*i.e.*, pour toute règle $R_i \in \mathcal{R}$, nous avons $R' \in \mathcal{C}(R_i)$). La confiance(R') $\geq max\{$confiance(R_i)$\}$ et support(R') $\geq max\{$support(R_i)$\}$.

IV)Inférence de règles basée sur la fermeture ($\mathcal{IRF}$)

Ce mécanisme d'inférence est basé sur le fait que le support d'un itemset fréquent I est égal à celui de sa fermeture $\omega(I)$ [33]. Ainsi, soit une règle $R : X \Rightarrow (Y - X)$, alors nous avons :

1. support($R : X \Rightarrow (Y - X)$) = support($\omega(Y)$) ;
2. confiance($R : X \Rightarrow (Y - X)$) = $\frac{support(\omega(Y))}{support(\omega(X))}$;
3. Une règle $R : X \Rightarrow (Y - X)$ est valide si et seulement si $R' : \omega(X) \Rightarrow (\omega(Y) - \omega(X))$ est valide.