Construction et exploration de résumés de grands ensembles de règles d'association

Marie Ndiaye[1,2], Cheikh T. Diop[2], Arnaud Giacometti[1], Patrick Marcel[1], Arnaud Soulet[1]

[1]LI - Université François Rabelais Tours
41000 Blois (France)
marie.ndiaye,arnaud.giacometti@univ-tours.fr
patick.marcel,arnaud.soulet@univ-tours.fr

[2]LANI - Université Gaston Berger de Saint-Louis
BP 234 Saint-Louis (Sénégal)
cdiop@ugb.sn

Le résumé est très utilisé pour représenter de grands ensembles de motifs, en particulier les ensembles de règles d'association. Généralement, les résumés de règles d'association proposés dans la littérature ne peuvent être présentés que sous forme de liste. De ce fait, il est difficile de les analyser. Dans ce travail, nous proposons une fonction qui construit des résumés de grands ensembles de règles qui peuvent être visualisés avec des tableaux croisés. Puis, nous définissons une mesure qui permet d'évaluer la qualité de ces résumés. Nous utilisons ensuite cette mesure dans un algorithme pour déterminer une solution approchée du résumé de qualité maximale dont la taille ne dépasse pas un seuil fixé. Par ailleurs, pour un ensemble de règles, on peut trouver plusieurs résumés intéressants d'où l'intérêt de naviguer entre eux. Enfin, les évaluations de notre algorithme effectuées sur des jeux de données réelles montrent la faisabilité de notre approche.

Mots clés : règles d'association, résumé, mesure de qualité de résumé.

1 Introduction

Les règles d'association sont des motifs introduits dans [Agrawal et al.]. Elles sont issues des algorithmes de data mining et sont utilisées pour découvrir des connaissances intéressantes à partir de données. Les algorithmes standards d'extraction produisent généralement une quantité très importante de règles. Il est alors difficile voire impossible d'identifier une connaissance pertinente à partir de ces ensembles de règles volumineux. Dans [Diop et al., 2008], nous avons étudié des techniques de visualisation usuelles et nous avons observé que leur efficacité est limitée lorsqu'il y a un nombre important de règles. En effet, les règles étant visualisées individuellement avec ces techniques, l'utilisateur est submergé quand il y en a beaucoup. D'autre part, plusieurs travaux ont été menés dans le but de réduire le nombre de règles extraites en supprimant les règles redondantes [Zaki, 2000a] ou peu intéressantes pour l'utilisateur [Bing et al., 1999], ou en introduisant des contraintes lors de l'extraction [Srikant et al., 1997]. Ces techniques peuvent diminuer considérablement le nombre de règles. Toutefois, la quantité reste souvent très importante. Pour interpréter plus facilement ces règles qui restent encore nombreuses, des méthodes de

résumé ont été proposées dans la littérature. Certains de ces travaux concernent le résumé d'ensembles de motifs [Yan et al., 2005, Chandola and Kumar, 2007, Afrati et al., 2004, Mielikäinen and Mannila, 2003, Jin et al., 2008]. D'autres s'intéressent en particulier au résumé d'ensembles de règles d'association [Ordonez et al., 2006].
Cependant, les résumés générés par les méthodes proposées dans ces travaux sont généralement difficiles à visualiser. En effet, la représentation ultérieure des résumés n'étant pas prise en compte lors de leur construction, il est difficile voire impossible de les visualiser autrement que sous forme de liste de motifs. Par ailleurs, pour un ensemble de règles donné, on peut trouver plusieurs résumés intéressants, d'où l'intérêt d'avoir la possibilité de naviguer entre eux. A notre connaissance, il n'existe pas de travail allant dans le sens de fournir un moyen de navigation entre des résumés.

La première contribution de ce papier est la proposition d'un format de résumé facile à visualiser avec les tableaux croisés et qui permet de définir des opérations intuitives de navigation entre résumés. La deuxième contribution est la proposition d'une mesure de qualité basée sur l'entropie de Shannon et permettant d'évaluer la représentativité des résumés. Enfin, la dernière contribution est la proposition d'un algorithme qui donne une solution approchée du résumé de qualité maximale et de taille inférieure à un seuil donné. Nous utilisons dans cet algorithme une recherche gloutonne en nous servant de certains opérateurs de navigation que nous avons définis. Notre algorithme est utilisé en post-traitement, i.e. il reçoit en entrée des règles extraites au préalable. Par ailleurs, il peut s'utiliser pour résumer aussi bien des bases ordinaires que des bases génériques de règles. Nos tests empiriques sur des bases génériques démontrent la faisabilité de notre approche qui fournit des résumés dans des délais raisonnables, même si le nombre de règles est très grand. Ils montrent également que les résumés de taille acceptable pour la visualisation ont une qualité intéressante.

Le reste de l'article est organisé comme suit. Quelques définitions et notations sont exposées dans la section 2. La section 3 est consacrée à la description des résumés que nous proposons. Dans la section 4, nous présentons une méthode de visualisation de ces résumés et des opérations élémentaires qui permettent de passer d'un résumé à un autre. Ensuite, nous proposons dans la section 5, une mesure pour évaluer la qualité des résumés. L'algorithme de construction de résumé est détaillé dans la section 6. Dans la section 7, nous présentons une évaluation de l'algorithme sur des données réelles. Un état de l'art sur les méthodes existantes de construction de résumé est présenté dans la section 8. Enfin, nous présentons une conclusion et des perspectives dans la section 9. Les preuves des propriétés énoncées dans le reste du papier sont détaillées en annexe.

2 Cadre général

Les données utilisées pour extraire les règles d'association sont issues de bases de données transactionnelles. Le tableau 1 présente une base de données qui contient des informations sur l'atterrissage d'engins spatiaux. Les données sont utilisées pour générer des règles permettant de déterminer les conditions sous lesquelles il est préférable d'utiliser un contrôle manuel ou automatique des engins. Chaque ligne correspond à un atterrissage sauf la première qui contient le nom des attributs observés. La première colonne contient les identifiants des enregistrements.

Tid	CONTROL	STABILITY	ERROR	SIGN	WIND	MAGNITUDE	VISIBILITY
0001	auto						no
0002	noauto	xstab					yes
0003	noauto	stab	LX				yes
0004	noauto	stab	XL				yes
0005	noauto	stab	MM	nn	tail		yes
0006	noauto					OutOfRange	yes
0007	auto	stab	SS			Low	yes
0008	auto	stab	SS			Medium	yes
0009	auto	stab	SS			Strong	yes
0010	auto	stab	MM	pp	head	Low	yes
0011	auto	stab	MM	pp	head	Medium	yes
0012	auto	stab	MM	pp	tail	Low	yes
0013	auto	stab	MM	pp	tail	Medium	yes
0014	noauto	stab	MM	pp	head	Strong	yes
0015	auto	stab	MM	pp	tail	Strong	yes

TAB. 1 – Base de données

2.1 Définitions et notations

Dans ce papier, on note $\mathcal{A}$ l'ensemble des attributs du domaine étudié. Soit A, un attribut de $\mathcal{A}$, on note $dom(A)$ le domaine de l'attribut A et $dom(\mathcal{A})$ le produit cartésien des domaines des attributs de $\mathcal{A}$, i.e. $dom(\mathcal{A}) = \times_{A \in \mathcal{A}} dom(A)$.

Exemple 1 *L'ensemble d'attributs du domaine étudié pour les données du tableau 1 est :* $\mathcal{A}$ *=*{*CONTROL, STABILITY, ERROR, SIGN, WIND, MAGNITUDE, VISIBILITY*}*. Le domaine de chaque attribut de* $\mathcal{A}$ *est présenté sur une ligne du tableau 2.*

Attribut	**Domaine**
CONTROL	{noauto, auto}
STABILITY	{stab, xstab}
ERROR	{XL, LX, MM, SS}
SIGN	{pp, nn}
WIND	{head, tail}
MAGNITUDE	{Low, Medium, Strong, OutOfRange}
VISIBILITY	{yes, no}

TAB. 2 – Attributs

Un item x défini sur $\mathcal{A}$ est un couple attribut-valeur noté $(A = a)$ avec $A \in \mathcal{A}$ et $a \in dom(A)$. Par la suite, $(A = a)$ est noté plus simplement a si aucune ambiguïté n'est possible, i.e. si les domaines des attributs de $\mathcal{A}$ sont disjoints.
Un itemset X défini sur $\mathcal{A}$ est un ensemble d'items définis sur $\mathcal{A}$. Soit un itemset $X = \{(A_1 = a_{i_1}), ..., (A_k = a_{i_k})\}$ défini sur $\mathcal{A}$, on appelle schéma de X noté $sch(X)$ l'ensemble d'attributs $\{A_1, ..., A_k\} \subseteq \mathcal{A}$.
Une règle d'association est une relation $X \Rightarrow Y$ où X et Y sont des itemsets et $X \cap Y = \emptyset$. X est appelé le corps et Y la tête de la règle. On appelle schéma de $X \Rightarrow Y$ noté $sch(X \Rightarrow Y)$ le couple $\langle sch(X), sch(Y) \rangle$ formé par le schéma de X et celui de Y.

Exemple 2 *Le tableau 3 présente des règles extraites à partir des données du tableau 1. La deuxième et la troisième colonne du tableau contiennent respectivement le corps et la tête des règles.*

Rule	Body	Head
r_1	{auto}	{stab}
r_2	{auto}	{stab,yes}
r_3	{auto}	{yes}
r_4	{stab}	{yes}
r_5	{stab}	{auto}
r_6	{stab}	{auto,yes}
r_7	{yes}	{stab}
r_8	{yes}	{auto,stab}
r_9	{yes}	{auto}

TAB. 3 – Ensemble de règles d'association

Une relation généralement utilisée pour construire des résumés de motifs est la relation de couverture. Dans ce papier, nous utilisons comme relation de couverture la relation suivante.

Définition 1 (Couverture) *Soient $r : X \Rightarrow Y$ et $r' : X' \Rightarrow Y'$ deux règles d'association. r est couverte par r' si r' est plus générale que r, i.e. $X' \subseteq X$ et $Y' \subseteq Y$. La relation r' « est plus générale que » r est notée $r' < r$.*

Exemple 3 *$r_1 : \{auto\} \Rightarrow \{stab\}$ couvre $r_2 : \{auto\} \Rightarrow \{stab, yes\}$ car le corps de r_1 est inclus dans celui de r_2, de même que la tête de r_1 est incluse dans celle de r_2.*

Cette définition de couverture est une adaptation aux règles de celle qui a été proposée dans [Chandola and Kumar, 2007] pour les itemsets. Elle est plus générale que la couverture utilisée dans [Ordonez et al., 2006]. En effet, les auteurs considèrent que r' couvre r seulement si $X' \subseteq X$ et $Y = Y'$ sachant que Y est constitué d'un seul item. Étant donnés un ensemble de règles R et une règle r, l'ensemble composé des règles de R couvertes par r est noté $cover(r, R)$.
La propriété suivante est une extension aux règles d'association d'une propriété sur les itemsets introduite dans [Zaki, 2000b].

Propriété 1 (Couverture) *Étant donnés deux règles $X_1 \Rightarrow Y_1$ et $X_2 \Rightarrow Y_2$, et un ensemble de règles R, $cover(X_1 \cup X_2 \Rightarrow Y_1 \cup Y_2, R) = cover(X_1 \Rightarrow Y_1, R) \cap cover(X_2 \Rightarrow Y_2, R)$.*

Elle signifie que si une règle peut être obtenue par union respective de la tête et du corps de deux autres règles, alors l'ensemble des règles qu'elle couvre est obtenu par intersection des ensembles de règles couvertes par ces deux règles.
La définition 2 est une adaptation aux règles d'association de la définition de résumé d'itemsets proposée dans [Chandola and Kumar, 2007] qui repose sur la relation de couverture.

Définition 2 (Résumé) *Soit R un ensemble de règles. Un résumé de R est un ensemble de règles S tel que (i) pour toute règle r de R, il existe une règle s de S telle que s couvre r, (ii) chaque règle de S couvre un sous-ensemble non vide de R.*

Par la suite, nous notons $\mathcal{S}_R$ l'ensemble de tous les résumés possibles de R. Notons que d'après cette définition, un résumé n'est pas forcément inclus dans l'ensemble qu'on résume. Par ailleurs, un résumé trivial d'un ensemble de règles est l'ensemble lui même. Cependant, il n'est pas intéressant car un des buts qu'on veut atteindre en résumant un ensemble de règles est d'obtenir un ensemble plus petit. Pour trouver des résumés intéressants, nous orientons notre recherche vers les résumés minimaux.

Définition 3 (Résumé minimal) *Soit R un ensemble de règles et S un résumé de R. S est dit minimal s'il n'existe pas d'ensemble de règles $S' \subset S$ tel que S' est un résumé de R.*

En d'autres termes, un résumé d'un ensemble de règles est minimal s'il n'est plus un résumé de cet ensemble lorsqu'on lui enlève ne serait-ce qu'une règle. Les résumés minimaux de R sont ses plus petits résumés possibles.

2.2 Formulation du problème

Le nombre de règles produites par les algorithmes d'extraction est souvent si grand que l'analyste rencontre des difficultés pour les explorer. Un moyen de lui faciliter la tâche est de lui présenter un ensemble de règles plus concis généralement appelé résumé. Dans ce papier, notre objectif est de trouver des résumés intéressants pour l'exploration de règles d'association. Ainsi, nous nous focalisons sur les trois problèmes suivants.

Résumés faciles à visualiser et navigables Le premier problème auquel nous nous intéressons est la génération de résumés minimaux pouvant être visualisés de manière compréhensible pour l'exploration de grands ensembles de règles d'association. De plus, on doit pouvoir naviguer intuitivement entre les résumés d'un même ensemble de règles. Nous proposons dans la section 3 des résumés de forme particulière permettant de répondre à nos besoins. L'ensemble de ces résumés est noté $\mathcal{S}_{R,\mathcal{A}}$. Ensuite, nous proposons dans la section 4 une visualisation de ces résumés avec les tableaux croisés et des opérateurs de navigation.

Mesure de qualité de résumé Lorsqu'on présente un résumé, il est essentiel de lui associer au moins une mesure de qualité sur laquelle l'utilisateur peut s'appuyer pour juger de sa pertinence. Une mesure de qualité est une fonction ϕ qui évalue la représentativité d'un résumé S par rapport à un ensemble de règles R. Elle dépend donc aussi bien du résumé que de l'ensemble de règles qu'on résume. Ainsi, le second problème traité dans ce papier est la définition d'une mesure de qualité intéressante pour les résumés de règles d'association. Nous proposons dans la section 5 une mesure de qualité basée sur l'entropie de Shannon.

Construction de résumé de bonne qualité Étant donné un ensemble de règles R, notre troisième objectif est de trouver un résumé intéressant pour initialiser l'exploration des résumés de R, i.e. un résumé S^* de qualité maximale sachant que sa taille est plus petite qu'un seuil fixé N. $\mathcal{S}_R$ est potentiellement très grand. Malgré la restriction de l'espace de recherche à $\mathcal{S}_{R,\mathcal{A}}$ qui est très petit devant $\mathcal{S}_R$, on peut encore se retrouver avec une quantité importante de résumés.

Dans ce travail, nous nous limitons à trouver une solution approchée de résumé optimal à partir de $\mathcal{S}_{R,\mathcal{A}}$.

$$S^* = \underset{S \in \mathcal{S}_{R,\mathcal{A}}, |S^*| \leq N}{\arg\max} \phi(R, S)$$

Nous proposons dans la section 6 un algorithme glouton de recherche de S^*.

3 Résumé minimal basé sur un schéma

Dans cette section, nous caractérisons notre espace de recherche en décrivant la composition des résumés qui nous intéressent. Nous proposons des résumés dont la forme des règles permet non seulement de les présenter autrement que sous forme de liste mais aussi de définir des opérations qui permettent de naviguer entre eux. On pourra ainsi produire des visualisations de résumés faciles à interpréter et donner aux utilisateurs un moyen de les explorer. Dans la définition 4, nous présentons une fonction qui construit un nouvel ensemble de règles à partir d'un ensemble de règles à résumer. Nous montrons par la suite que l'ensemble de règles obtenu est un résumé minimal de l'ensemble à partir duquel il est construit.

Les notations de base suivantes nous permettront de caractériser formellement notre espace de recherche. Considérons l'ensemble d'attributs du domaine $\mathcal{A}$. Soit *null* une constante n'appartenant à aucun des domaines $dom(A)$, $(A \in \mathcal{A})$. On note $dom(A)^+$ le domaine de A auquel on ajoute la constante *null*, i.e. $dom(A)^+ = dom(A) \cup \{null\}$. Enfin, on note $dom(\mathcal{A})^+$ le produit cartésien des ensembles $dom(A)^+$ $(A \in \mathcal{A})$, i.e. $dom(\mathcal{A})^+ = \times_{A \in \mathcal{A}} dom(A)^+$. D'autre part, on appelle extension d'un itemset X noté X^+ l'ensemble d'items défini par : $X^+ = X \cup \{(A = null) \mid A \in \mathcal{A} \setminus sch(X)\}$. Intuitivement, on ajoute à X tous les items $(A = null)$ tels que A est un attribut de $\mathcal{A}$ n'apparaissant pas dans $sch(X)$. Cela permet de représenter tous les itemsets selon le même schéma $\mathcal{A}$.

Définition 4 (Constructeur de résumé minimal basé sur un schéma) *La fonction de construction de résumé minimal σ associe à un ensemble de règles R et un schéma de règles $\langle \mathcal{B}, \mathcal{H} \rangle$ l'ensemble S des règles $X \Rightarrow Y$ tels que $X \in dom(\mathcal{B})^+$, $Y \in dom(\mathcal{H})^+$ et $cover(X \Rightarrow Y, R) \neq \emptyset$. Par la suite, on dit que S est basé sur le schéma $\langle \mathcal{B}, \mathcal{H} \rangle$ qui est commun à toutes ses règles.*

$$\sigma : \mathcal{P}(\mathcal{R}) \times (\mathcal{P}(\mathcal{A}))^2 \rightarrow \mathcal{P}(\mathcal{R})$$
$$(R, \langle \mathcal{B}, \mathcal{H} \rangle) \rightarrow S$$
$$où \quad S = \{X \rightarrow Y \mid X \in dom(\mathcal{B})^+ \wedge Y \in dom(\mathcal{H})^+ \wedge cover(X \rightarrow Y, R) \neq \emptyset\}$$

Propriété 2 (Résumé) *Soit un ensemble de règles R et un schéma de règles $\langle \mathcal{B}, \mathcal{H} \rangle$. L'ensemble de règles $S = \sigma(R, \langle \mathcal{B}, \mathcal{H} \rangle)$ est un résumé de R.*

La particularité de ces résumés est que leurs règles ont le même schéma. Ce schéma présente un avantage pour la visualisation et la navigation dans le sens où il constitue une structure supplémentaire commune aux règles. Il peut ainsi être utilisé pour la construction de visualisations de résumés et pour la définition d'opérateurs de navigation.

Notons que R et S ne sont pas définis à partir du même langage, i.e. les règles de S sont constituées d'extensions d'itemsets contrairement à celles de R qui sont formées d'itemsets standards. Par ailleurs, la construction de résumé s'effectue de manière naïve après l'extraction des règles. Pour déterminer le résumé S de R basé sur le schéma $\langle \mathcal{B}, \mathcal{H} \rangle$, on construit d'abord les règles candidates en effectuant toutes les combinaisons possibles entre les itemsets de $dom(\mathcal{B})^+$ et ceux de $dom(\mathcal{H})^+$. Ensuite, on construit $cover(s, R)$ pour chaque candidat s. Si $cover(s, R)$ n'est pas vide, alors la règle s est sélectionnée. Les règles ainsi retenues forment le résumé S.

Maintenant, nous allons étudier les propriétés des résumés basés sur un schéma.

Propriété 3 (Partition) *Soit un ensemble de règles R et un schéma de règles $\langle \mathcal{B}, \mathcal{H} \rangle$. Étant donné le résumé $S = \sigma(R, \langle \mathcal{B}, \mathcal{H} \rangle)$, les ensembles $cover(s, R)$, $s \in S$ forment une partition de R notée $\mathcal{P}_S(R)$.*

Cette propriété montre que chaque règle de l'ensemble qu'on résume est couverte par une seule et unique règle du résumé. Elle permet de garantir à l'utilisateur que les résumés qu'on lui présente ne comportent pas de redondance. De plus, elle permet d'accroître la lisibilité et la compréhension des résumés lors de leur visualisation.

Propriété 4 (Minimalité) *Soit un ensemble de règles R et un schéma de règles $\langle \mathcal{B}, \mathcal{H} \rangle$. Le résumé $S = \sigma(R, \langle \mathcal{B}, \mathcal{H} \rangle)$ est minimal.*

Cette propriété nous permet d'assurer qu'on construit les plus petits résumés possibles. Par conséquent, on réduit au maximum le nombre de règles à présenter à l'utilisateur, on rend ainsi le résumé plus facile à interpréter.

Exemple 4 *Considérons l'ensemble R constitué des règles du tableau 3 et le schéma de règles $\langle$\{VISIBILITY,WIND\},\{CONTROL\}$\rangle$. Rappelons que $dom(VISIBILITY)^+ = \{yes, no, null\}$, $dom(CONTROL)^+ = \{auto, noauto, null\}$ et $dom(WIND)^+ = \{head, tail, null\}$. Les règles de schéma $\langle$\{VISIBILITY\},\{CONTROL\}$\rangle$ sont obtenues en faisant toutes les combinaisons possibles des valeurs de $dom($\{VISIBILITY,WIND\}$)^+$ pour le corps et de $dom($\{CONTROL\}$)^+$ pour la tête. Cependant, seules quatre d'entre elles couvrent au moins une règle de R. Ces quatre règles constituent un résumé de R noté $S = \sigma(R, \langle \{VISIBILITY, WIND\}, \{CONTROL\} \rangle)$. Elles sont présentées dans la première colonne du tableau 4. Dans la deuxième colonne du tableau, on trouve les règles couvertes par chaque règle de S.*

De plus, nous constatons que S est bien un résumé minimal. En effet, les sous-ensembles de R couverts par les règles de S forment une partition de R (propriété 3). Si on retire ne serait-ce qu'une règle de S, alors les règles qui restent ne forment plus un résumé de R puisqu'il y aurait des règles de R qui ne seraient pas couvertes.

Étant donné un ensemble R dont les règles sont extraites à partir d'un schéma de relation $\mathcal{A}$, nous nous restreignons à un sous-ensemble de $\mathcal{S}_R$ qui est l'ensemble des résumés minimaux $\mathcal{S}_{R,\mathcal{A}} = \{\sigma(R, \langle \mathcal{B}, \mathcal{H} \rangle) \mid \mathcal{B} \subseteq \mathcal{A} \wedge \mathcal{H} \subseteq \mathcal{A}\}$. Nous montrons dans la section 4 que ces résumés peuvent être visualisés facilement avec des tableaux croisés et qu'on peut définir sur eux des opérations de navigation.

Règle du résumé	**Règles couvertes**
$s_1 : \{VISIBILITY = null, WIND = null\} \Rightarrow \{CONTROL = null\}$	$\{r_1, r_2, r_3, r_4\}$
$s_2 : \{VISIBILITY = null, WIND = null\} \Rightarrow \{CONTROL = auto\}$	$\{r_5, r_6\}$
$s_3 : \{VISIBILITY = yes, WIND = null\} \Rightarrow \{CONTROL = null\}$	$\{r_7\}$
$s_4 : \{VISIBILITY = yes, WIND = null\} \Rightarrow \{CONTROL = auto\}$	$\{r_8, r_9\}$

TAB. 4 – Résumé minimal basé sur le schéma ⟨{VISIBILITY,WIND},{CONTROL}⟩

4 Visualisation et navigation

Nous présentons dans cette section une visualisation des résumés proposés dans la section 3 sous la forme de tableau croisé. Ensuite, nous introduisons des opérations qui permettent de naviguer entre ces résumés.

Les résumés introduits dans la section 3 sont faciles à visualiser. En effet, avec le schéma de leurs règles, on peut les représenter sous forme de tableau croisé comme l'illustre la figure 1. Les

			TETE	
			CONTROL= *null*	CONTROL= *auto*
CORPS	VISIBILITY= *null*	WIND= *null*	4	2
	VISIBILITY= *yes*	WIND= *null*	1	2

FIG. 1 – Visualisation du résumé du tableau 4 avec un tableau croisé

itemsets de $dom(\mathcal{B})^+$ sont imbriqués en ligne et ceux de $dom(\mathcal{H})^+$ sont imbriqués en colonne. La cellule correspondant à deux itemsets $X \in dom(\mathcal{B})^+$ et $Y \in dom(\mathcal{H})^+$ représente la règle $X \Rightarrow Y$. Elle contient des informations sur cette dernière. On peut visualiser dans les cellules des mesures globales associées aux groupes de règles $cover(X \Rightarrow Y)$, en particulier des agrégations des mesures d'intérêt des règles couvertes. Des mesures de qualité individuelles associées aux règles du résumé peuvent également être représentées. Sur la figure 1, chaque cellule contient une mesure agrégée qui est le nombre de règles couvertes par la règle correspondant à la cellule.
Cette visualisation est facile à comprendre par sa simplicité et elle permet d'explorer de manière intuitive de grands ensembles de règles d'association via leurs résumés. Le fait que le schéma des règles du résumé soit unique permet de faciliter la navigation entre résumés. En effet, passer d'un résumé à un autre revient simplement à modifier le schéma du premier résumé en ajoutant ou en supprimant des attributs de $\mathcal{B}$ ou de $\mathcal{H}$. Dans ce papier, nous considérons que la navigation entre résumés correspond à l'application d'opérateurs au schéma des règles des résumés. Un opérateur de navigation est une fonction qui modifie les composantes du schéma des règles d'un résumé. Nous distinguons quatre opérateurs élémentaires : $AddToBody$, $AddToHead$, $DeleteFromBody$ et $DeleteFromHead$. Ils permettent respectivement d'ajouter des attributs à $\mathcal{B}$ et $\mathcal{H}$ et de supprimer des attributs de ces ensembles.
Soit A un attribut de $\mathcal{A}$. Les opérateurs sont formellement définis comme suit.

$$\begin{aligned} AddToHead(\langle \mathcal{B}, \mathcal{H} \rangle, A) &= \langle \mathcal{B}, \mathcal{H} \cup \{A\} \rangle \\ AddToBody(\langle \mathcal{B}, \mathcal{H} \rangle, A) &= \langle \mathcal{B} \cup \{A\}, \mathcal{H} \rangle \\ DeleteFromHead(\langle \mathcal{B}, \mathcal{H} \rangle, A) &= \langle \mathcal{B}, \mathcal{H} \backslash \{A\} \rangle \\ DeleteFromBody(\langle \mathcal{B}, \mathcal{H} \rangle, A) &= \langle \mathcal{B} \backslash \{A\}, \mathcal{H} \rangle \end{aligned}$$

La figure 2 montre une utilisation concrète de l'opérateur *AddToBody*.

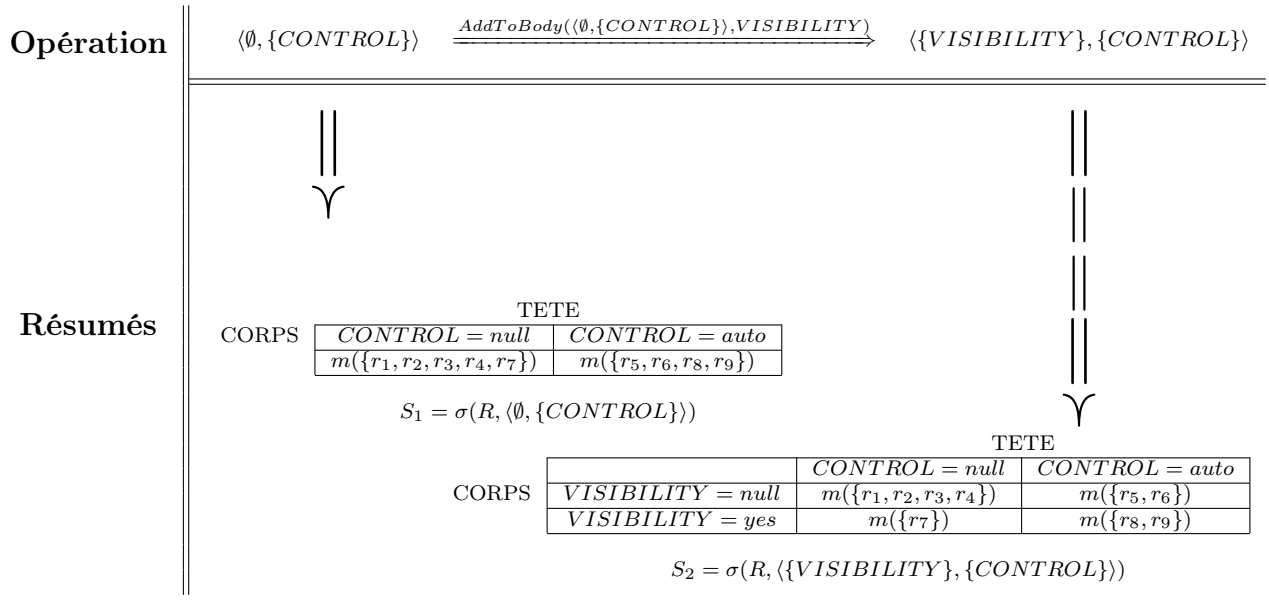

FIG. 2 – Navigation entre deux résumés

Le passage de S_1 à S_2 se fait en deux étapes : d'abord la construction des candidats avec le nouveau schéma $\langle\{VISIBILITY\}, \{CONTROL\}\rangle$, ensuite la sélection parmi les candidats, des règles devant constituer le résumé S_2. La construction des candidats est effectuée selon la méthode décrite après la propriété 2. Pour sélectionner les règles du résumé, nous construisons $\mathcal{P}_{S_2}(R)$ en utilisant $\mathcal{P}_{S_1}(R) = \{\{r_1, r_2, r_3, r_4, r_7\}, \{r_5, r_6, r_8, r_9\}\}$ qui, rappelons-le, est constituée des $cover(s, R)$, $s \in S_1$. $\mathcal{P}_{S_2}(R)$ est obtenue en partitionnant chaque ensemble de $\mathcal{P}_{S_1}(R)$ suivant les valeurs de $dom(VISIBILITY)^+$ qui sont *yes*, *no* et *null* (voir lemme 1 en annexe). Ainsi, nous découpons $\{r_1, r_2, r_3, r_4, r_7\}$ en $\{r_1, r_2, r_3, r_4\}$ pour la valeur *null* et $\{r_7\}$ pour la valeur *yes*, puis $\{r_5, r_6, r_8, r_9\}$ en $\{r_5, r_6\}$ pour la valeur *null* et $\{r_8, r_9\}$ pour la valeur *yes*. Dans les deux découpages, nous n'avons pas de partie associée à la valeur *no* car il n'existe pas de règle avec cette valeurs dans le corps. Donc les règles retenues sont celles qui couvrent les quatre nouvelles parties.

La visualisation de résumés est certes très intéressante mais puisque qu'il en existe plusieurs pour un ensemble de règles, il est aussi important de pouvoir identifier ceux qui sont intéressants. Nous introduisons dans la section 5 une nouvelle mesure de qualité qui permet à un utilisateur d'évaluer la représentativité d'un résumé lors de sa visualisation.

5 Mesure de qualité de résumé

Dans notre approche, un ensemble de règles est bien représenté par un résumé si les règles couvertes par chaque règle du résumé sont assez similaires, i.e. elles ont sensiblement la même tête et le même corps. Nous définissons dans cette section une mesure de qualité basée sur

l'entropie conditionnelle de Shannon [Shannon, 1948] qui permet d'évaluer cette similarité.
L'entropie conditionnelle de Shannon évalue l'homogénéité dans une population suivant une variable conditionnellement à une deuxième variable. Soient V_A et V_B deux variables aléatoires respectivement de n et m états possibles. L'entropie conditionnelle de V_A sachant V_B, notée $I(V_A|V_B)$, est définie par :

$$I(V_A|V_B) = \sum_{\substack{i\in\{1,\ldots,n\}\\ j\in\{1,\ldots,m\}}} p(V_A = a_i, V_B = b_j) \ln(p(V_A = a_i|V_B = b_j))$$

Considérons une règle s du résumé $S = \sigma(R, \langle \mathcal{B}, \mathcal{H}\rangle)$. Par construction, toutes les règles de $cover(s, R)$ ont la même valeur pour les attributs de $\mathcal{H}$ et de $\mathcal{B}$. Cependant nous n'avons aucune information concernant les valeurs prises pour les autres attributs de $\mathcal{A}$. Considérons un attribut $A \in \mathcal{A}\backslash(\mathcal{B} \cup \mathcal{H})$. La valeur prise pour A dans les règles de $cover(s, R)$ peut être la même ou différente. S'il s'agit de la même valeur, alors on dit que les règles de $cover(s, R)$ sont homogènes par rapport à A. Par contre s'il y a plusieurs valeurs prises, alors il y a un désordre engendré par la différence de ces valeurs.
Dans notre contexte, la population est l'ensemble de règles R, la première variable est un attribut $A \in \mathcal{A}$ et la seconde variable est le résumé S. L'homogénéité dans R suivant A, étant donné le résumé S, est évaluée par l'entropie de A conditionnellement à S calculée à partir de R.
Soit V_A la variable aléatoire discrète définie de R vers $dom(A)$ qui associe à une règle $X \Rightarrow Y \in R$, la valeur $a \in dom(A)^+$ si $(A = a) \in (X \cup Y)^+$. V_A est bien une fonction car on ne peut pas avoir dans une règle plus d'une valeur pour le même attribut. $\{V_A = a\}$ représente l'évènement « l'item $(A = a)$ appartient à la tête ou au corps d'une règle $X \Rightarrow Y$ de R », i.e $\{X \Rightarrow Y \in R \mid (A = a) \in (X \cup Y)^+\}$. Soit V_S la variable aléatoire discrète définie de R vers S qui associe à une règle $X \Rightarrow Y \in R$, la règle $s \in S$ telle que s couvre $X \Rightarrow Y$. Puisque les $cover(s, R)$ forment une partition de R (voir la propriété 3), alors chaque règle de R est couverte par une seule règle de S. $\{V_S = s\}$ correspond à l'évènement « s couvre une règle $X \Rightarrow Y$ de R », i.e $\{X \Rightarrow Y \in R \mid X \Rightarrow Y \in cover(s, R)\}$. La mesure de qualité de résumé est formellement définie comme suit.

Définition 5 (Qualité de résumé : homogénéité) *Soit R un ensemble de règles extraites à partir d'une relation de schéma $\mathcal{A}$ et S un résumé de R. L'homogénéité de S relativement à R est la moyenne des entropies conditionnelles des V_A, $A \in \mathcal{A}$ sachant V_S.*

$$\phi(R, S) = \frac{1}{|\mathcal{A}|} \sum_{A\in\mathcal{A}} I(V_A|V_S)$$

ϕ mesure l'homogénéité globale d'un résumé. Plus celle-ci augmente, plus la valeur retournée par ϕ sera élevée. Cette valeur est négative ou nulle. Elle vaut 0 si R est parfaitement homogène, i.e. si dans chaque groupe $cover(s, R)$, les règles contiennent les mêmes valeurs pour tous les attributs de $\mathcal{A}$.

Exemple 5 *Considérons le résumé construit dans l'exemple 4. Pour calculer sa qualité, nous utilisons les données du tableau suivant. Ce tableau présente le nombre de règles de $cover(s_i, R)$,*

Règle	Nbr couvertes	Attributs								
		CONTROL			STABILITY			VISIBILITY		
		auto	noauto	null	stab	xstab	null	yes	no	null
s_1	4	3	0	1	3	0	1	3	0	1
s_2	2	2	0	0	2	0	0	1	0	1
s_3	1	0	0	1	1	0	0	1	0	0
s_4	2	2	0	0	1	0	1	2	0	0
S	9	6	0	3	7	0	2	5	0	4

$i \in \{1, ..., 4\}$ qui contiennent chaque valeur d'attribut présent dans les règles du tableau 3. L'entropie conditionnelle est nulle pour tous les autres attributs qui n'apparaissent pas dans les règles.

$$\phi(R,S) = \frac{I(V_{CONTROL}|V_S) + I(V_{STABILITY}|V_S) + I(V_{VISIBILITY}|V_S)}{|\mathcal{A}|}$$

$$I(V_{CONTROL}|V_S) = \left[\frac{3}{9}\log(\frac{3}{4}) + \frac{1}{9}\log(\frac{1}{4})\right] + \left[\frac{2}{9}\log(\frac{2}{2})\right] + \left[\frac{1}{9}\log(\frac{1}{1})\right] + \left[\frac{2}{9}\log(\frac{2}{2})\right] = -0.25$$

$$I(V_{STABILITY}|V_S) = \left[\frac{3}{9}\log(\frac{3}{4}) + \frac{1}{9}\log(\frac{1}{4})\right] + \left[\frac{2}{9}\log(\frac{2}{2})\right] + \left[\frac{1}{9}\log(\frac{1}{1})\right] + \left[\frac{1}{9}\log(\frac{1}{2}) * 2\right] = -0.40$$

$$I(V_{VISIBILITY}|V_S) = \left[\frac{3}{9}\log(\frac{3}{4}) + \frac{1}{9}\log(\frac{1}{4})\right] + \left[\frac{1}{9}\log(\frac{1}{2}) * 2\right] + \left[\frac{1}{9}\log(\frac{1}{1})\right] + \left[\frac{2}{9}\log(\frac{2}{2})\right] = -0.40$$

D'où $\phi(R,S) = -\frac{2*0.40+0.25}{7} = -0.15$

Propriété 5 *Soient S_1, S_2 deux résumés de R dont les schémas de règles respectifs sont $\langle \mathcal{B}_1, \mathcal{H}_1 \rangle$ et $\langle \mathcal{B}_2, \mathcal{H}_2 \rangle$. Si S est un résumé de R basé sur le schéma $\langle \mathcal{B}_1 \cup \mathcal{B}_2, \mathcal{H}_1 \cup \mathcal{H}_2 \rangle$, alors $\phi(R,S) \geq \phi(R,S_i)$, $i \in \{1,2\}$.*

D'après la propriété 5, plus on ajoute des attributs dans $\mathcal{B}$ et $\mathcal{H}$, plus la qualité du résumé obtenu augmente. Il est donc facile de voir que le résumé dont le schéma est $\langle \emptyset, \emptyset \rangle$, est le résumé de plus mauvaise qualité et, que le résumé de schéma $\langle \mathcal{B}, \mathcal{H} \rangle$ tel que $\mathcal{B} = \mathcal{H} = \mathcal{A}$ est de qualité maximale, i.e. $\phi(R,S) = 0$.

La mesure de qualité est un indicateur permettant d'évaluer la pertinence des résumés. Cependant, s'il y a beaucoup de résumés qui sont proposés à l'utilisateur, ce dernier ne pourra pas les explorer tous. Nous proposons dans la section 6 une méthode de construction qui fournit un résumé intéressant qui peut être utilisé pour initialiser la navigation.

6 Construction de résumés

Dans cette section, notre objectif est de construire un premier résumé minimal intéressant à partir duquel l'utilisateur peut commencer la navigation.

La construction d'un bon résumé minimal repose essentiellement sur le choix du schéma de ses règles. Comme nous l'avons souligné à la fin de la section 5, les meilleurs résumés sont obtenus en mettant tout les attributs de $\mathcal{A}$ dans $\mathcal{B}$ et dans $\mathcal{H}$. Cependant, ces résumés ne sont pas intéressants car ils sont trop détaillés, i.e. ils ont une grande taille. Par conséquent, nous devons trouver le schéma $\langle \mathcal{B}, \mathcal{H} \rangle$ qui fournit un résumé de taille réduite et qui maximise la mesure de qualité ϕ.

Ainsi, le problème de construction de résumé peut être reformulé comme suit. Soit R un ensemble de règles extraites à partir d'une relation de schéma $\mathcal{A}$. Étant données la mesure de qualité ϕ et

une taille N : trouver un schéma $\langle \mathcal{B}, \mathcal{H} \rangle$ qui fournit un résumé de taille plus petit que N et de qualité maximale. La taille de l'espace de recherche $\mathcal{S}_{R,\mathcal{A}}$ est exponentielle en fonction du nombre d'attributs de $\mathcal{A}$. En effet, elle correspond à $2^{2\times|\mathcal{A}|}$. Son exploration devient coûteuse si A est grand. Nous proposons donc une solution approchée du problème avec l'algorithme glouton 1 en utilisant les opérateurs $AddToHead$ et $AddToBody$ présentées dans la section 3.

Algorithm 1 Algorithme glouton de construction de $\mathcal{H}$ et $\mathcal{B}$

Entrée : R {Ensemble de règles}, $\mathcal{A}$ {Ensemble d'attributs} et N {Seuil pour la taille du résumé}
Sorties : {Un schéma de résumé minimal}
Utilise : σ {Fonction de calcul de résumé}, ϕ {fonction de calcul de qualité de résumé},$AddToHead$ et $AddToBody${Operateurs d'ajout d'attribut}

1: $\mathcal{H}_0 = \emptyset$, $\mathcal{B}_0 = \emptyset$, $i = 0$ {Initialisations}
2: **répéter**
3: $i = i + 1$
4: $\mathcal{H}_i = \mathcal{H}_{i-1}$
5: $\mathcal{B}_i = \mathcal{B}_{i-1}$
6: **pour tout** $\mathcal{O} \in \{AddToHead, AddToBody\}$ **faire**
7: **pour tout** $A \in \mathcal{A}$ **faire**
8: $\langle \mathcal{H}, \mathcal{B} \rangle = \mathcal{O}(\langle \mathcal{H}_{i-1}, \mathcal{B}_{i-1} \rangle, A)$
9: **si** $\langle \mathcal{H}, \mathcal{B} \rangle \neq \langle \mathcal{H}_{i-1}, \mathcal{B}_{i-1} \rangle$ **alors**
10: **si** $\phi(R, \sigma(R, \langle \mathcal{B}, \mathcal{H} \rangle)) > \phi(R, \sigma(R, \langle \mathcal{B}_i, \mathcal{H}_i \rangle))$ **alors**
11: $\mathcal{H}_i = \mathcal{H}$
12: $\mathcal{B}_i = \mathcal{B}$
13: **finsi**
14: **finsi**
15: **fin pour**
16: **fin pour**
17: **jusqu'à** $|\sigma(R, \langle \mathcal{H}_i, \mathcal{B}_i \rangle)| > N$ ou $|\mathcal{H}_{i-1}| = |\mathcal{B}_{i-1}| = |\mathcal{A}|$
18: **retourner** $\langle \mathcal{H}_{i-1}, \mathcal{B}_{i-1} \rangle$

L'algorithme consiste à choisir a chaque étape i un attribut de $\mathcal{A}$ qui sera ajouté à $\mathcal{B}_{i-1}$ ou $\mathcal{H}_{i-1}$ pour obtenir $\mathcal{B}_i$ et $\mathcal{H}_i$ (ligne 8). Les ensembles d'attributs sont vides à la première étape. L'attribut retenu à l'étape i est choisi tel que le résumé de schéma $\langle \mathcal{B}_i, \mathcal{H}_i \rangle$ ait la plus grande qualité par rapport aux résumés obtenus en ajoutant un des autres attributs de $\mathcal{A}$ (lignes 6 à 16). On arrête d'ajouter des attributs lorsque la taille du résumé de schéma $\langle \mathcal{B}_i, \mathcal{H}_i \rangle$ dépasse le seuil fixé N ou quand les attributs de $\mathcal{A}$ sont tous dans $\mathcal{B}_i$ et $\mathcal{H}_i$ (ligne 17). Le schéma retourné $\langle \mathcal{B}_{i-1}, \mathcal{H}_{i-1} \rangle$ est celui qui est obtenu à l'avant dernière étape (ligne 18). Le résumé $\sigma(R, \langle \mathcal{B}_{i-1}, \mathcal{H}_{i-1} \rangle)$ est une solution approchée du résumé optimal S^*.

Nous utilisons uniquement les opérateurs $AddToHead$ et $AddToBody$ car la propriété 5 nous garantit qu'en ajoutant des attributs dans $\mathcal{H}$ ou $\mathcal{B}$, on augmente la qualité du résumé. L'implantation de l'algorithme nécessite le calcul des partitions associées aux résumés. A l'étape i, on calcule la partition associée à chaque résumé possible en utilisant la partition associée au résumé retenu à l'étape $i-1$. Nous optimisons ainsi le temps de calcul.

7 Expérimentations

L'objectif de cette section est d'évaluer la faisabilité de notre approche et son comportement. D'une part, nous étudions l'efficacité de notre algorithme en nous intéressant au temps d'exécution. D'autre part, nous observons l'évolution de la mesure d'homogénéité sur des données réelles.

Protocole expérimental Nous avons implémenté l'algorithme 1 en java, en utilisant la définition 5 de la mesure d'homogénéité. Pour les expérimentations, nous utilisons des versions discrétisées des jeux de données `mushroom`, `australian`, `vehicle` et `zoo` issus de l'UCI Machine Learning repository [1]. Nous utilisons des bases génériques de règles, extraites à partir de ces jeux de données avec l'algorithme CHARM [2] qui est décrit dans [Zaki and Hsiao, 2002]. Le tableau 5 indique le nombre de règles extraites à partir de ces jeux de données en fonction de la variation du support minimal $minsup$ avec une confiance minimale $minconf = 50\%$. Par ailleurs, toutes les expériences ont été réalisées sur un ordinateur Intel core duo 2GHz comportant 2 Go de mémoire vive sous Windows Vista. Plus précisément, nos analyses reposent sur deux expériences dont les protocoles sont les suivants :

- expérience 1 : le nombre de règles $|R|$ varie entre 0 et 30000 et la taille maximale du résumé est fixée à $N = 50$.
- expérience 2 : $minconf = 50\%$, $1 \leq N \leq 100$, $minsup = 25\%$ pour `mushroom` et `zoo` et $minsup = 15\%$ pour `australian` et `vehicle`.

mushroom : $|\mathcal{A}| = 23$, $minconf = 50\%$

minsup(%)	50	45	40	35	30	25	20	15	10
$\|R\|$	110	227	390	838	1732	2915	6681	11629	29442

vehicle : $|\mathcal{A}| = 19$, $minconf = 50\%$

minsup(%)	50	45	40	35	30	25	20	15	10	8
$\|R\|$	0	0	0	4	52	339	1066	3899	13890	318731

australian : $|\mathcal{A}| = 15$, $minconf = 50\%$

minsup(%)	50	45	40	35	30	25	20	15	10	5
$\|R\|$	9	23	62	124	247	486	1019	2437	7573	39060

zoo : $|\mathcal{A}| = 17$, $minconf = 50\%$

Support(%)	50	45	40	35	30	25	20	15	10
$\|R\|$	300	569	957	1864	3283	5382	8446	13253	20583

TAB. 5 – Nombre de règles générées en fonction du seuil de support

Temps d'exécution La figure 3a reporte le temps d'exécution selon le protocole de l'expérience 1. Nous remarquons que le temps d'exécution évolue linéairement en fonction du nombre de règles quelque soit la base. Nous observons également que la pente des courbes est déterminée par le nombre d'attributs des jeux de données. En effet, la courbe de `mushroom` a la plus grande pente car elle contient le plus grand nombre d'attributs (23 attributs). Elle est suivie par celle de `vehicle` avec 19 attributs, ensuite par celle de `zoo` avec 17 attributs et, enfin par celle de `australian` avec 15 attributs. Par ailleurs, nous voyons que le temps mis pour proposer un résumé n'excède pas 200 secondes même avec un ensemble dont la taille avoisine 30000 règles.

La figure 3b montre l'évolution du temps d'exécution en suivant le protocole de l'expérience 2. Nous observons que pour toutes les courbes, le temps d'exécution augmente de façon sous-linéaire en fonction de N. En effet, chaque attribut ajouté avec une opération $\mathcal{O}$ (cf. ligne 8 de l'algorithme 1) n'est plus testé avec cette même opération dans les étapes ultérieures (cf. ligne 9 de l'algorithme 1).

[1] `http://mlearn.ics.uci.edu/MLRepository.html` et `http://users.info.unicaen.fr/~frioult/uci`

[2] `http://http://www.cs.rpi.edu/~zaki/software/`

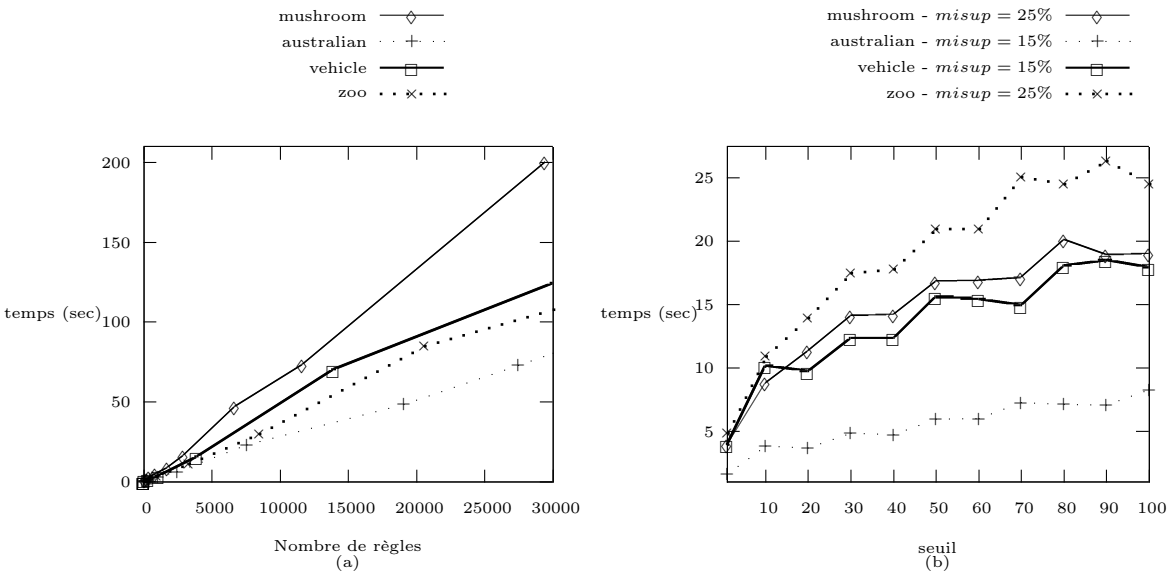

FIG. 3 – Temps d'exécution

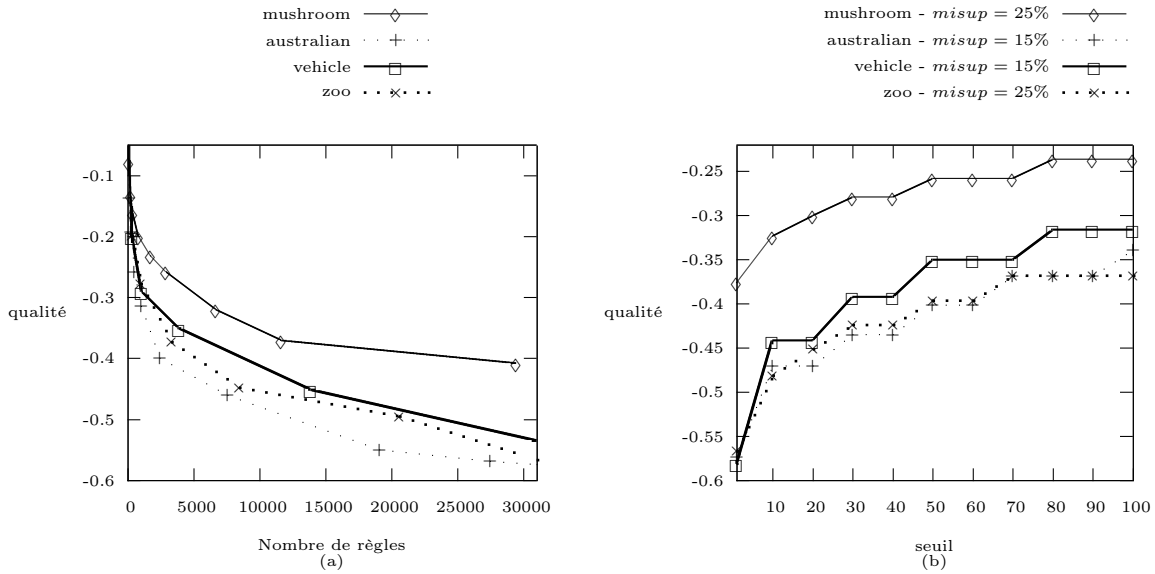

FIG. 4 – Qualité de résumé

Qualité La figure 4a présente l'évolution de l'homogénéité des résumés en appliquant le protocole de l'expérience 1. Nous observons que la qualité diminue au fur et à mesure que le nombre de règles augmente. Ce résultat était prévisible car plus le nombre de règles augmente (par diminution du support), plus on a des règles spécifiques qui s'ajoutent. Cette augmentation de règles spécifique a pour conséquence directe la diminution de l'homogénéité. Cependant, nous remarquons que cette décroissance est mieux que linéaire. D'autre part, nous observons que `mushroom` présente toujours la meilleure qualité, cela s'explique par le fait que ses attributs sont très corrélés.

La figure 4b reporte la variation de l'homogénéité des résumés en appliquant le protocole de l'expérience 2. Nous observons que la qualité augmente de façon logarithmique lorsqu'on fait croître N. Nous remarquons également que pour toutes les bases, cette augmentation est plus significative entre $N = 1$ et $N = 50$. Donc il est moins intéressant de fixer N au delà de 50 qui est d'ailleurs une taille raisonnable pour la visualisation.

De manière plus générale, ces expérimentations démontrent la faisabilité de notre approche qui fournit des résumés dans des délais raisonnables même si le nombre règles à résumer est très grand (au voisinage de 30000 règles). Elles montrent également que des résumés de taille acceptable pour la visualisation ont une qualité satisfaisante.

8 Travaux relatifs

La construction de résumés d'ensembles de motifs a été abordée selon plusieurs approches. Le tableau 6 présente une synthèse de méthodes de résumé proches de la notre.

Dans tous les papiers cités dans le tableau, les motifs considérés sont des itemsets, excepté dans [Ordonez et al., 2006] où les auteurs s'intéressent aux résumés d'ensembles de règles d'association qui concluent sur un item. La méthode de résumé proposée repose sur la couverture entre règles. Dans leur approche, une règle $X' \rightarrow y'$ couvre une autre règle $X \rightarrow y$ si $X' \subseteq X$ et $y' = y$. Les résumés proposés couvrent tout l'ensemble de règles. Toutefois, il n'y a pas de mesure de qualité de résumé définie.

Dans Afrati et al. [2004], Yan et al. [2005], Jin et al. [2008], Mielikäinen and Mannila [2003], la construction d'un résumé est considérée comme un problème de recherche de K motifs qui approximent au mieux un ensemble de motifs ou les mesures d'intérêt de ses motifs. La méthode proposée dans Mielikäinen and Mannila [2003] présente une certaine particularité. En effet, elle consiste à ordonner un ensemble de n motifs de sorte que pour tout $k < n$, l'ensemble de ses k premiers motifs soit un résumé qui approxime au mieux les $n - k$ motifs qui restent. De ce fait, le résumé de taille $k + 1$ est obtenu en ajoutant un motif au résumé de taille k alors que dans les autres travaux, on construit un nouveau résumé de taille k+1 qui peut être très différent de celui de taille k. Les mesures proposées dans Yan et al. [2005], Jin et al. [2008], Mielikäinen and Mannila [2003] évaluent l'erreur de restauration, i.e. l'erreur générée lorsqu'on approxime les motifs de l'ensemble ou leur mesure d'intérêt à partir du résumé. Par contre, la mesure utilisée dans Afrati et al. [2004] est la taille du sous-ensemble de motifs couverts par au moins un motif du résumé.

Deux mesures de qualité sont proposées dans [Chandola and Kumar, 2007] pour caractériser les

résumés : le gain de compacité qui évalue le taux de compression de l'ensemble de départ et la perte d'information qui évalue la représentativité du résumé par rapport à l'ensemble qu'il résume. Cette représentativité est calculée en quantifiant le nombre d'items présents dans chaque itemset de l'ensemble initial et absent de l'itemset qui le représente dans le résumé. La construction d'un résumé est considérée comme un problème de double optimisation : la maximisation du gain de compacité et la minimisation de la perte d'information.

Référence	[Ordonez et al., 2006]	[Chandola and Kumar, 2007]	[Afrati et al., 2004]	[Yan et al., 2005]	[Jin et al., 2008]	[Mielikäinen and Mannila, 2003]	Notre méthode
Motifs	règles	itemsets	itemsets	itemsets	itemsets	itemsets	règles
Approximation	non	non	ensemble de motifs	support des motifs	support des motifs	motifs	non
Couverture	approchée	exacte	approchée	approchée	exacte	approchée	exacte
Mesure	non	gain de compacité, perte d'information	nombre de motifs couverts	erreur de restauration	erreur de restauration	perte d'information	Entropie
Visualisation	non	non	non	non	non	non	oui
Navigation	non	non	non	non	non	non	oui

TAB. 6 – Méthodes de résumé de motifs

Les résumés que nous proposons couvrent totalement les ensembles de règles contrairement aux résumés proposés dans [Ordonez et al., 2006, Afrati et al., 2004, Yan et al., 2005, Mielikäinen and Mannila, 2003] qui couvrent approximativement les ensembles de règles. Nous nous intéressons à la représentativité des résumés comme les auteurs de [Chandola and Kumar, 2007], mais la mesure que nous proposons évalue plutôt la similarité entre les motifs couverts par les mêmes motifs du résumé. De manière générale, le but final de la construction de résumé est de faciliter l'analyse des motifs. Cette analyse passe donc par la visualisation des résumés qui doit être facile à interpréter. Cependant, contrairement à nous, les auteurs des travaux cités ci-dessus ne tiennent pas compte de la visualisation ultérieure des résumés lors de leur construction. Par ailleurs, comme on peut le voir dans [Mielikäinen and Mannila, 2003], on peut trouver plusieurs résumés intéressants pour un ensemble de règles donné, d'où l'intérêt de pouvoir naviguer entre eux. Il n'y a pas, à notre connaissance, de méthode de navigation proposée dans les travaux existants.

9 Conclusion et perspectives

Nous avons proposé une fonction qui fournit des résumés de grands ensembles de règles qui peuvent être présentés de manière compréhensible à l'utilisateur et qui permettent de définir des opérateurs pour une navigation intuitive entre les résumés. Par ailleurs, nous avons défini une mesure de qualité de résumé basée sur l'entropie et qui permet d'évaluer la représentativité de ces résumés. Enfin, nous avons proposé un algorithme qui construit un résumé de qualité maximale dont la taille ne dépasse pas un seuil fixé. Nos expérimentations démontrent la faisabilité de notre approche qui fournit des résumés dans des délais raisonnables même si le nombre règles à résumer est très grand. Elles montrent également que les résumés de taille acceptable pour la

visualisation ont une qualité intéressante. Ce travail est un premier pas dans la visualisation et la navigation de résumés de règles d'association reposant sur des méthodes OLAP.

Nous projetons de poursuivre ce travail en définissant de nouveaux opérateurs de navigation tels qu'un opérateur de zoom qui permet de se focaliser sur une partie de l'ensemble de règles, ou encore des opérateurs qui permettent de changer de niveau dans une hiérarchie des domaines des attributs. Ces opérateurs pourront être directement utilisés dans notre algorithme. Enfin, nous envisageons de proposer un outil d'exploration de grands ensembles de règles d'association basé sur la visualisation de résumés et, dont les interactions seront définies à partir des opérateurs de navigation.

Références

Foto Afrati, Aristides Gionis, and Heikki Mannila. Approximating a collection of frequent sets. In *KDD '04 : Proceedings of the tenth ACM SIGKDD*, pages 12–19, 2004. ISBN 1-58113-888-1.

Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD*, pages 207–216, FebruaryJune–FebruaryAugust˜ .

Liu Bing, Wynne Hsu, and Yiming Ma. Pruning and summarizing the discovered associations. In *KDD '99 : Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 125–134, 1999. ISBN 1-58113-143-7.

Varun Chandola and Vipin Kumar. Summarization – compressing data into an informative representation. *Knowl. Inf. Syst.*, 12(3) :355–378, 2007. ISSN 0219-1377.

Cheikh T. Diop, Arnaud Giacometti, Patrick Marcel, and Marie Ndiaye. Visualisation interactive de grands ensembles de règles d'association. In *8èmes journées francophones EGC : Atelier Visualisation et Extraction de Connaissances*, pages 31–51, 2008.

Ruoming Jin, Muad Abu-Ata, Yang Xiang, and Ning Ruan. Effective and efficient itemset pattern summarization : regression-based approaches. In *KDD '08 : Proceeding of the 14th ACM SIGKDD*, pages 399–407, 2008. ISBN 978-1-60558-193-4.

Taneli Mielikäinen and Heikki Mannila. The pattern ordering problem. In *Proceedings of the 7th European Conference on Principles of Data Mining and Knowledge Discovery, Lecture Notes in Artificial Intelligence*, pages 327–338. Springer-Verlag, 2003.

Carlos Ordonez, Norberto Ezquerra, and Cesar A. Santana. Constraining and summarizing association rules in medical data. *Knowl. Inf. Syst.*, 9(3) :259–283, 2006. ISSN 0219-1377.

Claude E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27 :379–423, 1948.

Ramakrishnan Srikant, Quoc Vu, and Rakesh Agrawal. Mining association rules with item constraints. In David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, editors, *KDD*, pages 67–73. AAAI Press, 14–17 1997.

Xifeng Yan, Hong Cheng, Jiawei Han, and Dong Xin. Summarizing itemset patterns : a profile-based approach. In *KDD '05*, pages 314–323, 2005. ISBN 1-59593-135-X.

Mohammed J. Zaki. Generating non-redundant association rules. In *KDD '00 : Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 34–43, 2000a. ISBN 1-58113-233-6.

Mohammed J. Zaki. Scalable algorithms for association mining. *IEEE Trans. on Knowl. and Data Eng.*, 12(3) :372–390, 2000b. ISSN 1041-4347.

Mohammed J. Zaki and Ching-Jui Hsiao. Charm : An efficient algorithm for closed itemset mining. In *SIAM 02*, pages 457–473, 2002.

Annexe

Preuve de la propriété 2 *Montrons que S satisfait aux propriétés (i) et (ii) de la définition 2.*

(i) Soit $X \Rightarrow Y$ une règle de R. Montrons qu'il existe une règle $X' \Rightarrow Y' \in S$ qui couvre $X \Rightarrow Y$. Considérons la règle $X^+ \Rightarrow Y^+$ qui est la représentation de $X \Rightarrow Y$ par l'extension de sa tête et de son corps. On a $sch(X^+) = sch(Y^+) = \mathcal{A}$. Soient $X' = \{(A = a) \in X^+ \mid A \in \mathcal{B}\}$ et $Y' = \{(A = a) \in X^+ \mid A \in \mathcal{H}\}$. La règle $X' \Rightarrow Y'$ a pour schéma $\langle \mathcal{B}, \mathcal{H} \rangle$. De plus, elle couvre au moins $X^+ \Rightarrow Y^+$ qui est l'extension de $X \Rightarrow Y$ contenu dans R. Donc, $X' \Rightarrow Y'$ appartient à S.

(ii) Cette propriété est triviale car d'après la définition de la fonction de résumé, toutes les règles présentes dans S couvrent au moins une règle de R.

Preuve de la propriété 3 *Montrons d'abord que $\underset{s \in S}{\cup} cover(s, R) = R$. Ensuite nous montrons que pour tout $s, s' \in S$, $cover(s, R) \cap cover(s', R) = \emptyset$.*
Pour tout $s \in S$, $cover(s, R) \subseteq R$, donc $\underset{s \in S}{\cup} cover(s, R) \subseteq R$. Soit r une règle de R. r est couverte par au moins une règle s de S, donc r appartient à $\underset{s \in S}{\cup} cover(s, R)$. Par conséquent, $\underset{s \in S}{\cup} cover(s, R) = R$.
Par ailleurs, considérons deux règles quelconques s et s' de S. Si une règle de R appartient à $cover(s, R) \cap cover(s', R)$, alors elle aurait dans sa tête ou son corps au moins deux items de même attribut et de valeur différentes puisque s et s' ont le même schéma, ce qui est impossible. Donc l'intersection de $cover(s, R)$ et $cover(s', R)$ est vide.

Preuve de la propriété 4 *Considérons un sous-ensemble quelconque de S noté S'. Montrons que S' n'est pas un résumé de R. Soit s une règle de $S \backslash S'$. Par définition, il existe au moins une règle $r \in R$ couverte par s. De plus, s est l'unique règle de S qui couvre r (voir la propriété 3). Ainsi, r n'est couverte par aucune règle de S'. Par conséquent, S' n'est pas un résumé de R car il ne satisfait pas à la propriété (i) de la définition 2.*

Lemme 1 *Soient un ensemble de règles R et trois résumés $S_1 = \sigma(R, \langle \mathcal{B}_1, \mathcal{H}_1 \rangle)$, $S_2 = \sigma(R, \langle \mathcal{B}_2, \mathcal{H}_2 \rangle)$ et $S = \sigma(R, \langle \mathcal{B}, \mathcal{H} \rangle)$ tel que $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ et $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$. La partition $\mathcal{P}_S(R)$ de R associée à S correspond à l'ensemble $\{R_1 \cap R_2 \mid R_1 \in \mathcal{P}_{S_1}(R) \wedge R_2 \in \mathcal{P}_{S_2}(R)\}$.*

Preuve du lemme 1 *Soit P un élément de $\mathcal{P}_S(R)$. Il existe une règle $X' \Rightarrow Y' \in S$ telle que $P = \{X \Rightarrow Y \in R | X' \subseteq X \wedge Y' \subseteq Y\}$. Puisque $X' \Rightarrow Y'$ a pour schéma $\langle \mathcal{B}_1 \cup \mathcal{B}_2, \mathcal{H}_1 \cup \mathcal{H}_2 \rangle$, alors il existe deux itemsets $X_1 \in dom(\mathcal{B}_1)^+$ et $X_2 \in dom(\mathcal{B}_2)^+$ tels que $X' = X_1 \cup X_2$ et deux itemsets $Y_1 \in dom(\mathcal{H}_1)^+$ et $Y_2 \in dom(\mathcal{H}_2)^+$ tels que $Y' = Y_1 \cup Y_2$. D'après la propriété 1, $P = cover(X_1 \Rightarrow Y_1, R) \cap cover(X_2 \Rightarrow Y_2, R)$, alors $cover(X_1 \Rightarrow Y_1, R)$ et $cover(X_2 \Rightarrow Y_2, R)$ sont non vides. Ainsi, $X_1 \Rightarrow Y_1 \in S_1$ et $X_2 \Rightarrow Y_2 \in S_2$. Par conséquent, $cover(X_1 \Rightarrow Y_1, R) \in \mathcal{P}_{S_1}$ et $cover(X_2 \Rightarrow Y_2, R) \in \mathcal{P}_{S_2}(R)$.*
Soit P un élément de $\{P_1 \cap P_2 \mid P_1 \in \mathcal{P}_{S_1}(R) \wedge P_2 \in \mathcal{P}_{S_2}(R)\}$. Il existe deux règles $X_1 \Rightarrow Y_1 \in S_1$ et $X_2 \Rightarrow Y_2 \in S_2$ telles que $P = cover(X_1 \Rightarrow Y_1, R) \cap cover(X_2 \Rightarrow Y_2, R)$. D'après la propriété 1, $P = cover(X_1 \cup X_2 \Rightarrow Y_1 \cup Y_2, R)$. Or, $X_1 \cup X_2 \Rightarrow Y_1 \cup Y_2$ a pour schéma $\langle \mathcal{B}_1 \cup \mathcal{B}_2, \mathcal{H}_1 \cup \mathcal{H}_2 \rangle$. De plus P est non vide. Donc $X_1 \cup X_2 \Rightarrow Y_1 \cup Y_2 \in S$, d'ou $P \in \mathcal{P}_S(R)$.

Preuve de la propriété 5 *D'après le lemme 1, chaque partie de $\mathcal{P}_S(R)$ est l'intersection d'une partie R_1 de $\mathcal{P}_{S_1}(R)$ et d'une partie R_2 de $\mathcal{P}_{S_2}(R)$. Ainsi, pour tout $s \in S$, il existe deux règles $s_1 \in S_1$ et $s_2 \in S_2$ telles que $\{V_S = s\} = \{V_{S_1} = s_1\} \cap \{V_{S_2} = s_2\}$. Inversement, pour tout $s_1 \in S_1$ et $s_2 \in S_2$, il existe une règle $s \in S$ telle que $\{V_{S_1} = s_1\} \cap \{V_{S_2} = s_2\} = \{V_S = s\}$.*
Par conséquent, $I(V_A \mid V_S) = I(V_A | V_{S_1}, V_{S_2}), A \in \mathcal{A}$

Or, $I(V_A|V_{S_1}, V_{S_2}) \geq I(V_A|V_{S_i})$, $i \in \{1, 2\}$ *(propriété de l'entropie)*
Donc $\frac{1}{|\mathcal{A}|} \sum_{A \in \mathcal{A}} I(V_A|V_{S_1}, V_{S_2}) \geq \frac{1}{|\mathcal{A}|} \sum_{A \in \mathcal{A}} I(V_A|V_{S_i})$, $i \in \{1, 2\}$