Fast-MGB : Nouvelle Base Générique Minimale de Règles Associatives

Chiraz Latiri*, Lamia Ben Ghezaiel**
Pr. Mohamed Ben Ahmed**

* Faculté des Sciences de Tunis
Département Informatique
Unité de recherche URPAH
Campus Universitaire El Manar, Tunis
Tunisie
chiraz.latiri@gnet.tn
**Ecole Nationale des Sciences de l'Informatique
Laboratoire RIADI-GDL
Campus Universitaire La Manouba, Tunis
Tunsisie
lamia.benghezaiel@riadi.rnu.tn
mohamed.benahmed@riadi.rnu.tn

Résumé. Le problème de l'exploitation des règles associatives est devenu primordial, puisque le nombre des règles associatives extraites des jeux de données réelles devient très élevé. Une solution possible consiste à ne dériver qu'une base générique de règles associatives. Cet ensemble de taille réduite permet de générer toutes les règles associatives via un système axiomatique adéquat. Dans cet article, nous proposons une nouvelle approche FAST-MGB qui permet de dériver, directement à partir du contexte d'extraction formel, une base générique minimale de règles associatives.

1 Introduction

Dans le cadre de ce travail, nous nous intéressons au problème d'extraction de règles associatives, initialement introduit par Agrawal et al. Agrawal et al. (1993). Plusieurs travaux basés sur l'analyse formelle des concepts (AFC) Ganter et Wille (1999), proposent des approches de sélection de règles associatives qui véhiculent le maximum de connaissances utiles. Ces approches reposent généralement sur l'extraction d'un sous-ensemble générique de toutes les règles associatives, appelé *base générique*, tout en satisfaisant certaines caractéristiques jugeant de sa qualité, mais qui dans la plupart des cas ne sont pas satisfaites dans leurs totalités Kryszkiewicz (2002).

Dans cet article, nous introduisons une nouvelle approche de génération d'une base minimale et générique (MGB) de règles associatives. L'originalité de cette approche est qu'elle est autonome : elle commence directement à partir du contexte d'extraction pour dériver une base générique minimale de règles associatives FAST-MGB.

2 Fondements mathématiques

Dans cette section, nous rappelons brièvement les notions mathématiques relatives à l'analyse formelle des concepts Ganter et Wille (1999).

2.1 Notions de bases

Contexte de fouille. Un contexte de fouille est un triplet $\mathcal{D} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ décrivant un ensemble fini $\mathcal{O}$ d'objets, un ensemble fini $\mathcal{I}$ d'attributs et une relation binaire $\mathcal{R}$. Chaque couple $(o, i) \in \mathcal{R}$, désigne que l'objet $o \in \mathcal{O}$, possède l'item $i \in \mathcal{I}$.

Pour $A \subseteq \mathcal{O}$, on définit $f(A) = \{d \mid \forall g, g \in A \Rightarrow (g, d) \in \mathcal{R}\}$; et pour $B \subseteq \mathcal{I}$, $h(B) = \{g \mid \forall d, d \in B \Rightarrow (g, d) \in R\}$.

L'opérateur de fermeture de la connexion de Galois Ganter et Wille (1999) est la composition des fonctions f et h, i.e. $(f \circ h)$.

Concept Réduit Fréquent. Soit $C \subseteq \mathcal{I}$ un ensemble d'items. C est appelé *Concept Réduit(CR)*, si et seulement s'il est égal à sa *fermeture*, i.e., $C = f \circ h(C)$. Il est dit *fréquent* (CRF) si $support(C) \geq$*minsupp*, où $support(C) = \frac{\|h(C)\|}{\|\mathcal{O}\|}$. L'ensemble des concepts réduits fréquents forme un semi-treillis appelé *treillis de l'Iceberg de Galois*.

Générateur minimal. Un itemset $g \subseteq \mathcal{I}$ est dit générateur (minimal) d'un concept réduit C si et seulement si $f \circ h(g) = C$ et $\nexists g' \subseteq g$ tel que $f \circ h(g') = C$ Bastide et al. (2000).

2.2 Les algorithmes d'extraction de bases génériques

Généralement, le nombre de *règles* associatives dérivées par un processus de fouille de données peut devenir très important surtout quant les mesures de fréquences deviennent assez faibles ou encore dans le cas de bases de données denses telles que les données textuelles.

Une solution possible à ce problème serait de se restreindre à l'extraction des *règles* strictement liées aux besoins de l'utilisateur, en d'autres termes, se limiter à une *base générique des règles associatives*, répondant à certains critères et à partir de laquelle les règles redondantes pourront être dérivées. Différentes approches de dérivation de bases génériques ont été proposées dans la littérature présentant chacune certaines limites. Nous distinguons quatre bases à savoir : la Base *représentative* (RB$_{Phan}$) Luong (2001), la Base des règles associatives *non redondantes* (MNR) Bastide et al. (2000), la Base générique pour les *règles représentatives* (RR) Kryszkiewicz (2002), et la Base Générique *Informative* (IGB) Gasmi et al. (2004). Une étude comparative des quatre bases est donnée dans Latiri et al. (2005).

Dans le cadre de notre travail, nous définissons une *base générique* comme suit :

Définition 1 *Une base générique est un ensemble de taille réduite de règles associatives ne contenant aucune règle redondante. Il existe trois critères pour évaluer une base générique quantitativement (en nombre de règles) et qualitativement (par la sémantique), à savoir :*

1. **Informativité** *: c'est la possibilité de déterminer avec exactitude le support et la confiance des règles redondantes dérivées à partir de la base générique, ce qui nécessite de garder la trace des concepts réduits fermés fréquents ou de leurs générateurs.*

2. **Compacité** *: c'est le fait d'avoir une base minimale, i.e. contenant le moins de règles possibles. Il faut trouver ainsi un compromis entre la compacité et l'informativité de ces bases.*
3. **Système axiomatique** *: c'est un ensemble de règles d'inférences permettant de dériver les règles redondantes. Un système axiomatique adéquat doit être complet (i.e. permet de dériver toutes les règles redondantes), et valide (i.e. toutes les règles redondantes dérivées doivent être valides par rapport à la valeur de minconf).*

Nous passons dans ce qui suit à la présentation de la nouvelle base générique FAST-MGB.

3 Une nouvelle base Générique minimale : FAST-MGB

Notre contribution consiste à introduire une *nouvelle base générique minimale des règles associatives non redondantes*, notée par FAST-MGB et ne contenant que des *règles implicatives* (i.e. prémisse différente de $\emptyset$). Cette approche assure la *compacité* et une *informativité partielle* avec un *système axiomatique complet et valide*. L'originalité de cette approche c'est qu'elle est autonome : elle commence directement à partir du contexte d'extraction pour dériver l'ensemble des concepts réduits fréquents, le treillis de l'iceberg de Galois et la base minimale générique de règles associatives FAST-MGB.

3.1 Génération de la base FAST-MGB

Dans le cadre de notre travail, nous considérons les règles associatives qui minimisent le nombre de termes dans la prémisse et qui maximisent le nombre de termes dans la conclusion.

3.1.1 Définition formelle

La base générique FAST-MGB est définie comme suit :

Définition 2 *Soit AR_k l'ensemble des règles associatives pouvant être extraites à partir d'un contexte d'extraction k. Une règle $R : X \rightarrow Y \in AR_k$ est redondante par rapport à $R_1 : X_1 \rightarrow Y_1$ si et seulement si les deux conditions suivantes sont vérifiées :*

1. $f \circ h(XY) = f \circ h(X_1Y_1)$
2. $X_1 \subseteq X \wedge Y \subset Y_1$

Dans la suite, nous définissons la base FAST-MGB comme suit :

Définition 3 *Soient,*

1. *$\mathcal{L}_c$: le treillis de l'iceberg de Galois, contenant tous les itemsets fermés fréquents pouvant être extraits à partir d'un contexte d'extraction k, et associés à leurs géné rateurs minimaux ainsi que leurs supports respectifs.*
2. *$\mathcal{S}$: l'ensemble des successeurs immédiats d'un itemset fermé fréquent c_i.*
3. *$\mathcal{G}_{c_i}$: l'ensemble des générateurs minimaux d'un itemset fermé fréquent c_i.*

$\mathcal{C}$	*le contexte d'extraction*
$E\_\mathcal{CRF}$	*L'ensemble des Concepts Réduits Fréquents.*
seuil-sup	*Seuil de support minimum requis pour q'un générateur soit retenu comme conclusion d'une règle*
minconf	*Seuil de confiance minimal exigé.*
$\mathcal{T}$	*Treillis de l'iceberg de Galois*

TAB. 1 – *Notations utilisées par l'algorithme* GEN-FAST-MGB

Formellement la base générique FAST-MGB *est définie comme suit :*

$$\text{FAST-MGB} = \left\{ \begin{array}{c} R : g \rightarrow c_i - g \mid g \in G_{c_i} \wedge c_i \in \mathcal{L}_c \wedge confiance(R) \geq minconf \\ \wedge \nexists\, s \in \mathcal{S} \mid \frac{support(s)}{support(g)} \geq minconf \end{array} \right\}$$

3.1.2 L'algorithme GEN-FAST-MGB

Dans ce qui suit, nous présentons l'algorithme de construction de la base FAST-MGB directement à partir du contexte d'extraction (voir l'algorithme 1 et le tableau 1).

Algorithme GEN-FAST-MGB
Entrée: $\mathcal{C}$: contexte d'extraction et les seuils *minsupp* et *minconf*
Sortie: FAST-MGB.
{/* Etape1*/}
$E\_\mathcal{CRF}$= GEN-CRF($\mathcal{C}$,*minsupp*)
{/*Etape2/*}
$\mathcal{T}$=GEN-TREILLIS($E\_\mathcal{CRF}$)
{/*Etape3*/}
FAST-MGB=GEN-REGLE($\mathcal{T}$,*minconf*)
Retourner(FAST-MGB)

Algorithm 1: L'algorithme GEN-FAST-MGB pour la dérivation de MGB directement à partir du contexte d'extraction

La génération de la base FAST-MGB s'effectue en 3 étapes, à savoir :

Étape1 : Générer l'ensemble des concepts réduits fréquents du treillis de l'iceberg de Galois enrichi par les générateurs minimaux L'algorithme GEN-CRF (voir algorithme 2) est itératif. Dans chaque itération k, il construit un ensemble de concepts formels réduits candidats (CRC_k) qui sera élagué ensuite, en respectant la contrainte de *minsupp*.

Étape 2 : Générer le treillis de l'iceberg de Galois enrichi par les générateurs minimaux La dérivation du treillis de l'iceberg de Galois, illustrée par l'algorithme 3 passe principalement par deux étapes, à savoir, *i)* La génération de la liste de tous les successeurs d'un concept donné et *ii)* À partir de la liste ainsi obtenue, ne retenir que les successeurs directement placés au-dessus du concept en question.

Algorithme GEN-CRF
Entrée: $\mathcal{C}$: contexte d'extraction et le seuil support minimal *minsupp*.
Sortie: $E\_\mathcal{CRF}$
$CRC_1 = \{1 - itemsets\}$
Pour tout (k=1 ;CRC_k.gen $\neq \emptyset$;k++) **Faire**
CRF_k= Gen-concepts(CRC_k)
CRC_{k+1} =Gen-next(CRF_k)
Retourner CRF=$\cup_k CRF_k$

Algorithm 2: L'algorithme GEN-CRF pour la génération des concepts réduits fréquents

Étape 3 : Générer la base FAST-MGB La base FAST-MGB constitue un ensemble réduit de règles implicatives. Le pseudo-code GEN-REGLE est donné par l'algorithme 4.

Algorithme GEN-TREILLIS
Entrée: : $E\_\mathcal{CRF}$ = {c_1,c_2,...,c_l}.
Sortie: $\mathcal{T}$: le treillis de l'icberg de Galois enrichi par les générateurs minimaux.
Pour tout $(k = 1; k \leq l; k++)$ **Faire**
Pour tout c$\in$ c_k **Faire**
l.pred = prédécesseur(c) ;
l.pred_direct = pred_Direct(c) ;
l.successeur=successeur(c) ;
$\mathcal{T} = \mathcal{T} \cup I$

Algorithm 3: L'algorithme GEN-TREILLIS pour la génération du treillis de l'iceberg de Galois

Algorithme GEN-REGLE
Entrée: $\mathcal{L}_c$:le treillis del 'icberg de Galois enrichi par les générateurs minimaux et minconf.
Sortie: $FAST - MGB$
Pour tout concept $c_i \in \mathcal{L}_c$ **Faire**
$G_{c_i} = \{$générateurs minimaux du concept c$'$|c$'$ $\subseteq c_i\}$
Pour tout générateur $g \in G_{C_i}$ **Faire**
Si $\nexists$ s $\in S_{c_i}$ tel que $\frac{support(s)}{support(g)} > minconf$ et S = {successeur immédiat de c_i} **alors**
R=R $\cup\, g \Rightarrow c_i - g$
G_{c_i} =$G_{c_i} - \{g' | g \subseteq g'\}$
Sinon
$G_{c_i} = G_{c_i} - g$
retourner (R) ;

Algorithm 4: L'algorithme GEN-REGLE pour la génération de la base de règles minimales

4 Conclusion et travaux en cours

Dans cet article, nous avons proposé une nouvelle approche d'extraction de base générique minimale de règles associatives entre termes directement à partir du contexte d'extraction. Nous proposons aussi d'étendre le champ d'application du Text Mining, moyennant la technique de découverte de règles associatives entre termes, aux textes en langue arabe. Nous avons

ainsi élaboré la phase la plus complexe du processus du fouille textuelle à savoir, le prétraitement linguistique. Pour ce faire, cette phase est subdivisée en deux étapes à savoir : *i)* L'analyse des textes ; *ii)* L'élaboration du contexte d'extraction textuel.

L'ensemble des travaux en cours s'articule sur deux axes à savoir, *i)* Finaliser la plateforme visuelle qui permet l'intérgration des différentes étapes pour la génération de la base générique minimale FAST-MGB et *ii)* Finaliser les tests de FAST-MGB sur les corpus textuels en langue arabe déjà pré-traités.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Records*, 207–216.

Bastide, Y., N. Pasquier, R. Taouil, L. Lakhal, et G. Stumme (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In *Proceedings of the International Conference DOOD'2000, Lecture Notes in Computer Sciences, Springer-verlag*, pp. 972–986.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis*. Edition Springer-Verlag, Heidelberg.

Gasmi, G., S. BenYahia, E. M. Nguifo, et Y. Slimani (2004). A new informative generic base of association rules. In *Proceedings of the 2nd Inl. Workshop on Concept Lattices and Applications (CLAŠ04), Ostrava, Czech Republic*, pp. 67–79.

Kryszkiewicz, M. (2002). Concise representations of association rules. In *Proceedings of Pattern Detection and Discovery, ESF Exploratory Workshop, London, UK*, pp. 92–109.

Latiri, C., W. Bellagha, et S. Benyahia (2005). VIE-MGB : A visual Interactive Exploration of Minimal Generic Basis of Association rules. In *Proceedings of the third International Conference on Concept Lattices and Their Applications (CLA 05), Olomouc, Czech Republic*, pp. 179–196.

Luong, V. P. (2001). Raisonnement sur les règles d'association. In *Dans les actes des* $17^{ème}$ *Journées Bases de Données Avancées BDA'2001, Agadir(Maroc), Cépaduès Edition*, pp. 299–310.

Summary

Mining association rules is an important task, even though the number of rules discovered is often huge. A possible solution to this problem, is to use the Formal Concept Analysis (FCA) as mathematical settings to restrict rules extraction to a generic basis of association rules. In this paper, we introduce a new minimal generic basis FAST-MGB of non-redundant association rules based on the augmented Iceberg Galois lattice.