Les itemsets essentiels fermés : une nouvelle représentation concise

Tarek Hamrouni*, Islem Denden*
Sadok Ben Yahia*, Engelbert Mephu Nguifo**, Yahya Slimani*

* Département des Sciences de l'Informatique
Faculté des Sciences de Tunis
Campus Universitaire 1060 Tunis, Tunisie
{tarek.hamrouni, sadok.benyahia, yahya.slimani}@fst.rnu.tn
** CRIL CNRS FRE 2499
Université d'Artois, IUT de Lens
Rue Jean Souvraz, SP-18
F-62307 Lens Cedex France
mephu@cril.univ-artois.fr

Résumé. Devant l'accroissement constant des grandes bases de données, plusieurs travaux de recherche en fouille de données s'orientent vers le développement de techniques de représentation compacte. Ces recherches se développent suivant deux axes complémentaires : l'extraction de bases génériques de règles d'association et l'extraction de représentations concises d'itemsets fréquents.

Dans ce papier, nous introduisons une nouvelle représentation concise exacte des itemsets fréquents. Elle se situe au croisement de chemins de deux autres représentations concises, à savoir les itemsets fermés et ceux dits essentiels. L'idée intuitive est de profiter du fait que tout opérateur de fermeture induit une fonction surjective. Dans ce contexte, nous introduisons un nouvel opérateur de fermeture permettant de calculer les fermetures des itemsets essentiels. Ceci a pour but d'avoir une représentation concise de taille réduite tout en permettant l'extraction des supports négatif et disjonctif d'un itemset en plus de son support conjonctif. Un nouvel algorithme appelé D-CLOSURE permettant d'extraire les itemsets essentiels fermés est aussi présenté. L'étude expérimentale que nous avons menée a permis de confirmer que la nouvelle approche présente un bon taux de compacité comparativement aux autres représentations concises exactes.

1 Introduction

L'apparition de la "fouille de connaissances" a été un tournant dans les intérêts prioritaires de la communauté de la fouille de données. En effet, les efforts ne sont plus seulement déployés dans la réduction des temps d'extraction des motifs fréquents mais de plus en plus de travaux s'intéressent à l'extraction d'une connaissance de meilleure qualité tout en préservant la vertu de la compacité. Dans ce registre, nous relevons les travaux visant l'extraction des représentations concises. Ainsi, parmi les représentations exactes les plus connues, nous citons

celles basées respectivement sur les itemsets fermés (Pasquier et al. (1999)), les itemsets non-dérivables (Calders et Goethals (2002)) et les itemsets essentiels (Casali et al. (2005)). Bien qu'offrant un taux de compacité intéressant, la représentation basée sur les itemsets essentiels souffre de son association avec la bordure positive afin de la rendre exacte.

Dans ce papier, nous introduisons un nouvel opérateur permettant d'extraire *les fermetures disjonctives* des itemsets essentiels fréquents. Nous obtenons ainsi les itemsets essentiels fermés qui forment une représentation concise exacte des itemsets fréquents. Deux particularités sont à mettre au crédit de cette nouvelle représentation : *(i)* La dérivation aisée des supports disjonctifs et négatifs des itemsets ; *(ii)* L'élimination de la bordure positive puisque l'ensemble des itemsets essentiels fermés constitue à lui seul une représentation concise exacte des itemsets fréquents. Les expérimentations que nous avons menées sur des bases benchmark ont montré que la nouvelle représentation concise présente un taux de compacité largement meilleur que les autres représentations concises. En particulier, nous arrivons même à réduire les représentations des bases éparses, là où les itemsets fermés ont échoué à le faire.

Le papier est organisé comme suit : dans la section 2, nous rappelons les concepts de base, et principalement ceux inhérents aux différents types de supports. Dans la section 3, nous passons en revue les principes de la représentation concise basée sur les itemsets essentiels. Dans la section 4, nous présentons les opérateurs disjonctifs ainsi que leurs propriétés. Dans la section 5, nous introduisons la définition formelle de notre représentation concise et qui sera suivie, dans la section 6, par la description de l'algorithme D-CLOSURE permettant l'extraction des itemsets essentiels fermés. Dans la section 7, nous présentons une étude expérimentale permettant de comparer la cardinalité de la représentation proposée avec celles de l'ensemble de tous les itemsets fréquents et des représentations basées sur les itemsets fermés fréquents et essentiels fréquents. La section 8 conclut le papier avec un rappel de notre contribution et des pistes de travaux futurs.

2 Concepts de base

Dans cette section, nous présentons les concepts de base qui seront utilisés par la suite.

Définition 1 (CONTEXTE FORMEL) *Un contexte formel (ou contexte d'extraction) est un triplet $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, décrivant deux ensembles finis $\mathcal{O}$ et $\mathcal{I}$ et une relation (d'incidence) binaire, $\mathcal{R}$, entre $\mathcal{O}$ et $\mathcal{I}$ tel que $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$. L'ensemble $\mathcal{O}$ est habituellement appelé ensemble d'objets (ou transactions) et $\mathcal{I}$ est appelé ensemble d'items (ou attributs). Chaque couple (o,i) $\in \mathcal{R}$ désigne que l'objet $o \in \mathcal{O}$ possède l'item $i \in \mathcal{I}$ (noté $o\mathcal{R}i$).*

Exemple 1 *Un exemple de contexte d'extraction $\mathcal{K}$ est présenté par la figure 1 avec $\mathcal{O} = \{1, 2, 3, 4, 5\}$ et $\mathcal{I}$=$\{a, b, c, d, e, f\}$.*

La définition suivante introduit les différents types de supports pouvant être associés à un itemset.

Définition 2 *(Casali et al. (2005))* (SUPPORTS D'UN ITEMSET) *Soit un contexte d'extraction $\mathcal{K}$ = ($\mathcal{O}$, $\mathcal{I}$, $\mathcal{R}$). Nous distinguons trois types de supports associés à un itemset I :*
*- **Support conjonctif :** Supp(I) =* $|\{o \in \mathcal{O} \mid (\forall\, i \in I, (o, i) \in \mathcal{R})\}|$
*- **Support disjonctif :** Supp($\vee$I) =* $|\{o \in \mathcal{O} \mid (\exists\, i \in I, (o, i) \in \mathcal{R})\}|$

	a	b	c	d	e	f
1	×	×	×	×		
2			×	×	×	
3	×	×			×	×
4	×	×	×	×	×	×
5			×	×		×

FIG. 1 – *Un exemple de contexte d'extraction.*

- Support négatif : *Supp(*$\neg I$*) =* $\mid \{o \in \mathcal{O} \mid (\forall\, i \in I, (o,i) \notin \mathcal{R})\} \mid$

Remarque 1 *Notons que Supp(*$\vee\emptyset$*) n'existe pas vu que l'ensemble vide ne contient aucun item.*

Exemple 2 *Considérons le contexte d'extraction de la figure 1. Les différents supports qui peuvent être associés à l'itemset* bc [1] *sont : Supp(*bc*) = 2, Supp(*$\vee bc$*) = 5, Supp(*$\neg bc$*) = 0.*

Le lemme suivant établit les différentes relations qui existent entre les différents supports d'un itemset I. Ces relations sont basées sur les identités d'inclusion-exclusion (Galambos et Simonelli (2000)).

Lemme 1 *(Galambos et Simonelli (2000))* (IDENTITÉS D'INCLUSION-EXCLUSION) *Les identités d'inclusion-exclusion établissent les liens qui existent entre le support conjonctif, le support disjonctif et le support négatif d'un itemset quelconque* I.

$$Supp(I) \;=\; \sum_{\substack{I_1 \subseteq I \\ I_1 \neq \emptyset}} (-1)^{\mid I_1 \mid - 1}\; Supp(\vee I_1)$$

$$Supp(\vee I) \;=\; \sum_{\substack{I_1 \subseteq I \\ I_1 \neq \emptyset}} (-1)^{\mid I_1 \mid - 1}\; Supp(I_1)$$

$$Supp(\neg I) \;=\; \mid \mathcal{O} \mid \;-\; Supp(\vee I)\ (\textit{La loi de De Morgan})$$

Ce lemme nous permet d'affirmer que la connaissance du support disjonctif (resp. conjonctif) de tous les sous-ensembles de I nous permet de calculer le support conjonctif (resp. disjonctif) de I. De même, la connaissance du support disjonctif (resp. négatif) de I nous permet de dériver directement le support négatif (resp. disjonctif) de I.

Exemple 3 *Considérons le contexte d'extraction de la figure 1. Nous allons montrer comment calculer les différents supports de l'itemset* bc *grâce aux identités d'inclusion-exclusion.*

- *Supp(*bc*) =* $(-1)^{|bc|\,-\,1}$ *Supp(*$\vee bc$*) +* $(-1)^{|b|\,-\,1}$ *Supp(*$\vee b$*) +* $(-1)^{|c|\,-\,1}$ *Supp(*$\vee c$*) = - Supp(*$\vee bc$*) + Supp(*$\vee b$*) + Supp(*$\vee c$*) = - 5 + 3 + 4 = 2.*
- *Supp(*$\vee bc$*) =* $(-1)^{|bc|\,-\,1}$ *Supp(*bc*) +* $(-1)^{|b|\,-\,1}$ *Supp(b) +* $(-1)^{|c|\,-\,1}$ *Supp(*c*) = - Supp(*bc*) + Supp(b) + Supp(*c*) = - 2 + 3 + 4 = 5.*
- *Supp(*$\neg bc$*) =* $|\mathcal{O}|$ *- Supp(*$\vee bc$*) = 5 - Supp(*$\vee bc$*) = 5 - 5 = 0.*

[1] Les ensembles d'items seront représentés sans séparateurs, *e.g.*, *bc* représente l'ensemble *{b, c}*.

3 Travaux antérieurs

La représentation concise que nous allons introduire est inspirée de deux notions complémentaires, à savoir les itemsets fermés fréquents (Pasquier et al. (1999)) et les itemsets essentiels fréquents (Casali et al. (2005)). Dans ce qui suit, nous allons nous focaliser sur les propriétés structurelles de la représentation concise basée sur les itemsets essentiels. Une étude des principales représentations concises restantes se trouve dans (Calders et al. (2006)).

La représentation concise exacte basée sur les itemsets essentiels possède deux avantages majeurs : d'une part, elle offre à l'utilisateur la possibilité de calculer les différents types de supports (*cf.* Définition 2) et d'autre part, elle présente un taux de compacité intéressant (Casali et al. (2005)).

Définition 3 *(Casali et al. (2005))* (ITEMSET ESSENTIEL FRÉQUENT) *Soit $\mathcal{K}$ = ($\mathcal{O}$, $\mathcal{I}$, $\mathcal{R}$) un contexte d'extraction et $I \subseteq \mathcal{I}$. I est un itemset* essentiel *si et seulement si Supp($\vee I$) $\neq \max\{$Supp($\vee I \backslash i$) $\mid i \in I\}$. Un itemset essentiel I est* fréquent *si Supp(I) $\geq$ minsup.*

Exemple 4 *Considérons le contexte d'extraction de la figure 1 pour minsup = 1. ab n'est pas un itemset essentiel puisque Supp($\vee ab$) = $\max\{$Supp($\vee a$), Supp($\vee b$)$\}$= Supp($\vee a$) = 3 alors que ac est un itemset essentiel puisque Supp($\vee ac$) $\neq \max\{$Supp($\vee a$), Supp($\vee c$)$\}$ (car $\max\{$Supp($\vee a$), Supp($\vee c$)$\}$ = Supp($\vee c$) = 4 et Supp($\vee ac$) = 5). De plus, ac est fréquent puisque Supp(ac) = 2 $\geq$ minsup.*

La proposition suivante affirme que l'ensemble des itemsets essentiels fréquents vérifie une propriété intéressante qui est le fait d'être un idéal d'ordre (Ganter et Wille (1999)).

Proposition 1 *(Casali et al. (2005)) L'ensemble des itemsets essentiels fréquents est un idéal d'ordre.*

Donc, si I est un itemset essentiel fréquent, alors tous ses sous-ensembles sont des itemsets essentiels fréquents. D'une manière duale, si I n'est pas un itemset essentiel fréquent alors tous ses sur-ensembles ne sont pas des itemsets essentiels fréquents. Cette propriété intéressante permet aux algorithmes d'extraction par niveau d'extraire d'une manière cet ensemble. Dans cet esprit, l'algorithme GLAE (Casali et al. (2005)), permettant d'extraire les itemsets essentiels fréquents, est une adaptation de l'algorithme pionnier d'extraction par niveau, à savoir APRIORI (Agrawal et Srikant (1994)).

Dans ce qui suit, nous désignerons par $\mathcal{IEF}_{\mathcal{K}}$ (resp. $\mathcal{IF}_{\mathcal{K}}$) l'ensemble des itemsets essentiels fréquents (resp. itemsets fréquents) qui peuvent être extraits à partir du contexte d'extraction $\mathcal{K}$. Le lemme suivant montre comment nous pouvons obtenir le support disjonctif d'un itemset fréquent à partir de l'ensemble $\mathcal{IEF}_{\mathcal{K}}$.

Lemme 2 *(Casali et al. (2005))* $\forall\ I \in \mathcal{IF}_{\mathcal{K}}$, *Supp($\vee I$) = max$\{$Supp($\vee I_1$) $\mid I_1 \subseteq I \wedge I_1 \in \mathcal{IEF}_{\mathcal{K}}\}$.*

Le théorème qui suit définit la représentation concise basée sur les itemsets essentiels fréquents.

Théorème 5 *(Casali et al. (2005)) L'ensemble $\mathcal{IEF}_{\mathcal{K}}$ des itemsets essentiels fréquents augmenté par l'ensemble BD^+ des itemsets maximaux fréquents est une représentation concise exacte de l'ensemble des itemsets fréquents.*

Ce théorème affirme que l'ensemble des itemsets essentiels fréquents ne forme une représentation concise exacte que si nous lui ajoutons l'ensemble BD^+. L'ajout de ce dernier aura pour conséquence l'augmentation de la cardinalité de la représentation basée sur les itemsets essentiels fréquents.

4 Opérateurs disjonctifs et leurs propriétés

L'idée de base de cette nouvelle représentation concise est d'appliquer un opérateur de fermeture sur les itemsets essentiels fréquents afin d'obtenir une représentation concise plus compacte. Cependant, cet opérateur est différent de celui appliqué dans le cas de la représentation concise basée sur les itemsets fermés fréquents (Pasquier et al. (1999)). En effet, les itemsets essentiels sont caractérisés via leurs supports *disjonctifs* et non pas ceux *conjonctifs*. Alors, nous avons besoin d'un nouvel opérateur de fermeture, que nous nommerons *opérateur de fermeture disjonctive*. L'intérêt de cette nouvelle représentation concise basée sur la fermeture disjonctive est double :
1. Avoir une représentation plus compacte que celle basée sur les essentiels fréquents. En effet, la fermeture disjonctive, comme tout opérateur de fermeture, est une fonction surjective. Le nombre des itemsets essentiels fermés sera dans tous les cas inférieur au nombre des itemsets essentiels fréquents. De plus, cette représentation concise va préserver l'avantage d'avoir l'information relative aux différents supports et non seulement le support conjonctif.
2. Éviter le besoin d'ajouter une information additionnelle afin de vérifier si un itemset est fréquent ou pas, telle que BD^+ dans le cas de la représentation concise basée sur les itemsets essentiels fréquents.

Afin de présenter la fermeture disjonctive, nous avons besoin de définir les applications correspondantes qui assurent le lien entre $\mathcal{P}(\mathcal{I})$ et $\mathcal{P}(\mathcal{O})$ et vice versa.

Définition 4 *Soit $\mathcal{K}$ = ($\mathcal{O}$, $\mathcal{I}$, $\mathcal{R}$) un contexte d'extraction. Les opérateurs assurant la connexion entre $\mathcal{P}(\mathcal{I})$ et $\mathcal{P}(\mathcal{O})$ sont les suivants :*

$$f_d : \mathcal{P}(\mathcal{O}) \rightarrow \mathcal{P}(\mathcal{I})$$
$$O \mapsto f_d(O) = \{i \in \mathcal{I} \mid [(\exists\, o \in \mathcal{O})\ ((o \in O) \wedge ((o,i) \in \mathcal{R}))] \wedge [(\forall\, o_1 \in \mathcal{O})\ ((o_1 \notin O) \Rightarrow ((o_1,i) \notin \mathcal{R}))]\}$$
$$g_d : \mathcal{P}(\mathcal{I}) \rightarrow \mathcal{P}(\mathcal{O})$$
$$I \mapsto g_d(I) = \{o \in \mathcal{O} \mid [(\exists\, i \in \mathcal{I})\ ((i \in I) \wedge ((o,i) \in \mathcal{R}))]\}$$

Exemple 6 *Si nous considérons le contexte représenté par la figure 1, nous avons alors : $f_d(\{4\}) = \emptyset$, $f_d(\{2, 3, 4, 5\}) = \{e, f\}$ et $g_d(\{a, c\}) = \{1, 2, 3, 4, 5\}$, $g_d(\{a, b\}) = \{1, 3, 4\}$.*

Après avoir énoncé les opérateurs de connexion, nous pouvons introduire leurs composées.

Définition 5 *(Hamrouni et al. (2006)) Soit $\mathcal{K}$ = ($\mathcal{O}$, $\mathcal{I}$, $\mathcal{R}$) un contexte d'extraction. Soit f_d et g_d les opérateurs définis dans la définition 4. Nous définissons les opérateurs composés résultants comme suit :*

$$
\begin{array}{rcl}
h_d = f_d \circ g_d : \mathcal{P}(\mathcal{I}) & \rightarrow & \mathcal{P}(\mathcal{I}) \\
I & \mapsto & h_d(I) = \{i \in \mathcal{I} \mid [(\exists\, o \in \mathcal{O})\ ((o,i) \in \mathcal{R})] \wedge \\
 & & [(\forall\, o_1 \in \mathcal{O})\ ((o_1,i) \in \mathcal{R}) \Rightarrow ((\exists\, i_1 \in \mathcal{I})\ ((i_1 \in I) \wedge \\
 & & ((o_1,i_1) \in \mathcal{R})))]\} \\
h'_d = g_d \circ f_d : \mathcal{P}(\mathcal{O}) & \rightarrow & \mathcal{P}(\mathcal{O}) \\
O & \mapsto & h'_d(O) = \{o \in \mathcal{O} \mid (\exists\, i \in \mathcal{I})\ ((o,i) \in \mathcal{R}) \wedge \\
 & & ((\forall\, o_1 \in \mathcal{O})\ (o_1 \notin O) \Rightarrow ((o_1,i) \notin \mathcal{R}))\}
\end{array}
$$

Exemple 7 *Considérons le contexte d'extraction de la figure 1. Nous avons :* $h_d(ac) = f_d \circ g_d(ac) = f_d(\{1, 2, 3, 4, 5\}) = abcdef$ *et* $h'_d(\{1, 3, 4\}) = g_d \circ f_d(\{1, 3, 4\}) = g_d(ab) = \{1, 3, 4\}$.

Maintenant, nous présentons les principales propriétés des opérateurs (composés) que nous avons introduits.

Proposition 2 *(Hamrouni et al. (2006)) Soit* $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ *un contexte d'extraction,* f_d *(resp.* g_d*) les opérateurs assurant le lien entre* $\mathcal{P}(\mathcal{I})$ *(resp.* $\mathcal{P}(\mathcal{O})$*) et* $\mathcal{P}(\mathcal{O})$ *(resp.* $\mathcal{P}(\mathcal{I})$*). Soit* I, I_1, $I_2 \in \mathcal{P}(\mathcal{I})$ *et* O, O_1, $O_2 \in \mathcal{P}(\mathcal{O})$*. Nous avons donc les propriétés suivantes :*

(1) $O_1 \subseteq O_2 \Rightarrow f_d(O_1) \subseteq f_d(O_2)$	(1') $I_1 \subseteq I_2 \Rightarrow g_d(I_1) \subseteq g_d(I_2)$
(2) $I \subseteq h_d(I)$	(2') $h'_d(O) \subseteq O$
(3) $I_1 \subseteq I_2 \Rightarrow h_d(I_1) \subseteq h_d(I_2)$	(3') $O_1 \subseteq O_2 \Rightarrow h'_d(O_1) \subseteq h'_d(O_2)$
(4) $f_d(O) = h'_d(f_d(O))$	(4') $g_d(I) = h_d(g_d(I))$
(5) $h_d(I) = h_d(h_d(I))$	(5') $h'_d(O) = h'_d(h'_d(O))$
(6) $g_d(I) \subseteq O \Leftrightarrow I \subseteq f_d(O)$	

h_d est un opérateur de fermeture puisqu'il vérifie les conditions requises. En effet, il est extensif (*cf.* Propriété (2)), isotone (*cf.* Propriété (3)) et idempotent (*cf.* Propriété (5)) [2]. Nous pouvons ainsi introduire la fermeture disjonctive (Hamrouni et al. (2006)).

Définition 6 (FERMETURE DISJONCTIVE) *Soit* $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ *un contexte d'extraction. L'opérateur de fermeture* h_d *est défini comme suit :*

$$
\begin{array}{rcl}
h_d : \mathcal{P}(\mathcal{I}) & \rightarrow & \mathcal{P}(\mathcal{I}) \\
I & \mapsto & h_d(I) = \{i \in \mathcal{I} \mid [(\exists\, o \in \mathcal{O})\ ((o,i) \in \mathcal{R})] \wedge \\
 & & [(\forall\, o_1 \in \mathcal{O})\ ((o_1,i) \in \mathcal{R}) \Rightarrow ((\exists\, i_1 \in \mathcal{I})\ ((i_1 \in I) \wedge \\
 & & ((o_1,i_1) \in \mathcal{R})))]\}
\end{array}
$$

Cette fermeture disjonctive sera à l'origine de la nouvelle représentation de l'ensemble des itemsets fréquents. Cette affirmation sera détaillée dans la section qui suit.

5 Une nouvelle représentation concise basée sur la fermeture disjonctive

Nous commençons par présenter la définition d'un itemset fermé disjonctif.

[2] h'_d n'est pas un opérateur de fermeture puisqu'il n'est pas extensif (*cf.* Propriété (2')). h'_d est dit opérateur d'ouverture (Ganter et Wille (1999)).

Définition 7 (ITEMSET FERMÉ DISJONCTIF) *Un itemset $I \subseteq \mathcal{I}$ est dit fermé disjonctif si et seulement si $h_d(I) = I$. Un itemset fermé disjonctif est l'ensemble maximal des items contenus uniquement dans l'ensemble des transactions dans lesquelles apparaît au moins un item de I et qui n'apparaissent nul part ailleurs.*

Exemple 8 *Étant donné le contexte d'extraction représenté par la figure 1, l'itemset ef est un itemset fermé disjonctif puisqu'il est égal à l'ensemble maximal d'items contenus uniquement dans l'ensemble des transactions qui contiennent au moins un item de ef, i.e., {2, 3, 4, 5}. D'où, $h_d(ef) = ef$. af n'est pas un itemset fermé disjonctif puisque b n'appartient pas à af alors qu'il apparaît uniquement dans l'ensemble des transactions où a ou f apparaissent. Ainsi, $h_d(af) = abf$.*

L'ensemble des itemsets fermés disjonctifs est défini comme suit :

Définition 8 (ENSEMBLE DES ITEMSETS FERMÉS DISJONCTIFS) *Soit $\mathcal{K}$ un contexte d'extraction et h_d l'opérateur de fermeture disjonctive. L'ensemble $\mathcal{IFD}_\mathcal{K}$ des itemsets fermés disjonctifs, extrait à partir d'un contexte $\mathcal{K}$, est défini comme suit : $\mathcal{IFD}_\mathcal{K} = \{I \subseteq \mathcal{I} \mid h_d(I) = I \wedge Supp(I) \geq minsup\}$.*

La proposition suivante nous permet d'établir la relation qui existe entre le plus petit itemset fermé disjonctif contenant I et $h_d(I)$.

Proposition 3 *(Hamrouni et al. (2006)) Le support disjonctif d'un itemset I est égal à celui du plus petit itemset fermé disjonctif le contenant.*

Par soucis de simplicité, nous allons désigner par la suite les itemsets fermés disjonctifs relatifs aux essentiels fréquents par les itemsets essentiels fermés tout court. Maintenant, nous pouvons introduire notre représentation concise exacte.

Théorème 9 *(Hamrouni et al. (2006)) L'ensemble $\mathcal{IFD}_\mathcal{K}$ des itemsets essentiels fermés est une représentation concise exacte de l'ensemble $\mathcal{IF}_\mathcal{K}$ des itemsets fréquents.*

Le lemme qui suit nous permet de garantir le fait que la cardinalité de notre représentation concise ne dépassera jamais celle des itemsets essentiels.

Lemme 3 *(Hamrouni et al. (2006)) La cardinalité de $\mathcal{IFD}_\mathcal{K}$ est au plus égale à $\mathcal{IEF}_\mathcal{K}$.*

Exemple 10 *Considérons le contexte d'extraction de la figure 1 pour minsup = 1. L'ensemble des itemsets essentiels fermés extrait à partir de ce contexte est présenté dans la table 1. Il est important de noter que pour 21 itemsets essentiels fréquents, nous avons seulement 8 itemsets essentiels fermés qui leur sont relatifs. Étant donné l'ensemble $\mathcal{IFD}_\mathcal{K}$, nous sommes en mesure de dériver le support conjonctif de chaque itemset fréquent.*

Supposons que nous voulions dériver le support conjonctif de l'itemset cef. Le plus petit itemset essentiel fermé contenant cef est $abcdef$. D'où, Supp($\vee cef$) = Supp($\vee abcdef$) = 5. Nous aurons aussi besoin des supports disjonctifs de tous les sous-ensembles de cef. D'après la proposition 3, nous avons Supp($\vee ce$) = Supp($\vee abcdef$) = 5, Supp($\vee cf$) = Supp($\vee abcdef$) = 5 et Supp($\vee c$) = Supp($\vee cd$) = 4. e, f et ef sont des itemsets essentiels fermés et nous pouvons immédiatement accéder à leurs supports disjonctifs qui sont égaux à 3, 3 et 4 respectivement. En appliquant les identités d'inclusion-exclusion, nous obtenons : Supp(cef) = Supp($\vee cef$) - Supp($\vee ce$) - Supp($\vee cf$) - Supp($\vee ef$) + Supp($\vee c$) + Supp($\vee e$) + Supp($\vee f$) = 5 - 5 - 5 - 4 + 4 + 3 + 3 = 1. D'où, Supp(cef) = 1.

Itemset essentiel fermé	Itemsets essentiels associés	Support disjonctif
e	e	3
f	f	3
ab	a, b	3
cd	c, d	4
ef	ef	4
abe	ae, be	4
abf	af, bf	4
$abcdef$	ac, ad, bc, bd, ce,	5
	cf, de, df, aef, bef	5

TAB. 1 – *L'ensemble $\mathcal{IFD}_{\mathcal{K}}$ pour minsup = 1.*

6 L'algorithme D-CLOSURE

Afin d'extraire l'ensemble $\mathcal{IFD}_{\mathcal{K}}$ des itemsets essentiels fermés, nous proposons un algorithme appelé D-CLOSURE (abréviation de Disjunctive CLOSURE), de type *"générer et tester"*. Afin d'expliciter les stratégies d'élagage utilisées, nous avons besoin d'introduire la proposition suivante.

Proposition 4 *Soient X et Y deux itemsets tels que $X \subseteq Y$ et $Y \subseteq h_d(X)$, alors $h_d(X) = h_d(Y)$ et $Supp(\vee X) = Supp(\vee Y)$.*

$\mathcal{K}$	Contexte d'extraction
C_i	L'ensemble des itemsets essentiels fréquents candidats de taille i.
L_i	L'ensemble des itemsets essentiels fréquents de taille i.
DC_i	L'ensemble des itemsets essentiels fermés, relatifs aux essentiels fréquents de taille i.
X_i	Itemset de taille i.
$X_i.h_d$	La fermeture disjonctive de l'itemset X_i.
$X_i.h_d^-$	L'ensemble des items qui ne doivent pas appartenir à $X_i.h_d$. Autrement dit, c'est l'ensemble des items qui sont apparus dans les transactions qui ne contiennent aucun item appartenant à X_i.
$X_i.Supp\_Conj$	Le support conjonctif de X_i.
$X_i.Supp\_Disj$	Le support disjonctif de X_i.

TAB. 2 – *Notations utilisées dans l'algorithme* D-CLOSURE.

Les stratégies d'élagage utilisées par D-CLOSURE sont les suivantes : *(i)* Un élagage par rapport à la fréquence conjonctive des itemsets essentiels candidats (*i.e.*, par rapport au seuil minimum de support *minsup*) ; *(ii)* Un élagage par rapport à l'idéal d'ordre, vérifié par l'ensemble des itemsets essentiels fréquents ; *(iii)* Un élagage par rapport à l'inclusion dans la fermeture de l'un des sous-ensembles (*cf.* Proposition 4).

Les notations utilisées par la suite sont données dans la table 2. Le pseudo-code de l'algorithme D-CLOSURE est donné par Algorithme 1. APRIORI_GEN est la procédure utilisée dans (Agrawal et Srikant (1994)) pour générer les candidats de taille $(i + 1)$ à partir des éléments retenus de taille i. Le pseudo-code de la procédure CALCUL_SUPPORTS_FERMETURE est donné

par Procédure 1. Cette procédure permet de calculer, en effectuant un seul parcours du contexte d'extraction, les supports conjonctifs et disjonctifs ainsi que la fermeture disjonctive de chaque itemset X_i appartenant à C_i.

Algorithme 1 : D-CLOSURE

Entrée : Le contexte d'extraction $\mathcal{K}$ et le seuil minimum de support *minsup*.
Sortie : $\cup_{j=1..i}\ DC_j$ qui est l'ensemble des essentiels fermés relatifs aux essentiels fréquents.

1 : $i := 1$;
2 : $C_1 := \mathcal{I}$;
3 : Tant que ($C_i \neq \emptyset$) faire
4 : CALCUL_SUPPORTS_FERMETURE($\mathcal{K}$, $minsup$, C_i, L_i, DC_i) ; /**L'élagage par rapport au seuil minimum $minsup$ s'effectue au sein de cette procédure**/
5 : C_{i+1} := APRIORI_GEN(L_i) ;
6 : $C_{i+1} := \{X_{i+1} \in C_{i+1} \mid \forall\ X_i \subset X_{i+1}, (X_i \in L_i) \wedge (X_{i+1} \not\subseteq X_i.h_d)\}$; /**Cette instruction permet d'effectuer l'élagage des itemsets essentiels candidats par rapport à l'idéal d'ordre et par rapport à l'inclusion du candidat dans la fermeture disjonctive de l'un de ses sous-ensembles immédiats**/
7 : $i := i + 1$;
8 : fin Tant que
9 : retourner $\cup_{j=1..i}\ DC_j$;

Procédure 1 : CALCUL_SUPPORTS_FERMETURE ($\mathcal{K}$, $minsup$, C_i, $var\ L_i$, $var\ DC_i$)

1 : Pour chaque (transaction $T \in \mathcal{K}$) faire
2 : Pour chaque (itemset $X_i \in C_i$) faire
3 : $\Omega := X_i \cap T$;
4 : Si ($\Omega = \emptyset$) alors
5 : $X_i.h_d^- := X_i.h_d^- \cup T$;
6 : Sinon
7 : $X_i.Supp\_Disj := X_i.Supp\_Disj + 1$;
8 : Si ($\Omega = X_i$) alors
9 : $X_i.Supp\_Conj := X_i.Supp\_Conj + 1$;
10 : fin Si
11 : fin Si
12 : fin Pour
13 : fin Pour
14 : Pour chaque (itemset $X_i \in C_i$) faire
15 : Si ($X_i.Supp\_Conj \geq minsup$) alors
16 : $L_i := L_i \cup \{X_i\}$;
17 : $X_i.h_d := \mathcal{I} \setminus X_i.h_d^-$;
18 : $DC_i := DC_i \cup \{(X_i.h_d, X_i.Supp\_Disj)\}$;
19 : fin Si
20 : fin Pour

7 Évaluation expérimentale

Dans cette section, nous présentons quelques résultats que nous avons obtenus en comparant la taille de notre représentation avec celle de l'ensemble total des itemsets fréquents

ainsi qu'avec celles des représentations basées sur les itemsets fermés fréquents et essentiels fréquents, respectivement. Les tests ont été effectués sur différentes bases benchmark [3] dont les caractéristiques sont résumées dans la table 3.

Base	Nombre d'items	Nombre de transactions	Taille moyenne des transactions
CONNECT	**129**	**67, 557**	**42**
MUSHROOM	**119**	**8, 124**	**23**
CHESS	**75**	**3, 196**	**37**
T10I4D100K	**1, 000**	**100, 000**	**10**

TAB. 3 – *Caractéristiques des bases de test.*

Les résultats obtenus sont présentés dans les tables 4-7. Au premier coup d'oeil, ces résultats nous permettent de déduire les affirmations suivantes :
1. Même pour des valeurs élevées de *minsup*, la cardinalité notre représentation concise $\mathcal{IFD}_{\mathcal{K}}$ est considérablement réduite par rapport à celle des itemsets fréquents $\mathcal{IF}_{\mathcal{K}}$ (*cf.*, $6^{ème}$ colonne et spécialement pour la base CONNECT).
2. Spécialement dans les bases CONNECT et CHESS, la cardinalité de $\mathcal{IFD}_{\mathcal{K}}$ est largement inférieure à celle des itemsets fermés fréquents $\mathcal{IFF}_{\mathcal{K}}$.
3. La quatrième colonne montre l'effet désastreux de l'addition de l'ensemble BD^+ par rapport au taux de compacité réalisé par l'ensemble des itemsets essentiels fréquents $\mathcal{IEF}_{\mathcal{K}}$. Heureusement, l'ensemble $\mathcal{IFD}_{\mathcal{K}}$ a surmonté cet handicap en excluant BD^+ de la représentation concise. Rappelons que la cardinalité de $\mathcal{IFD}_{\mathcal{K}}$ ne peut jamais dépasser celle de $\mathcal{IEF}_{\mathcal{K}}$.
4. Pour la base éparse T10I4D100K, nous remarquons que nous arrivons quand même à obtenir un taux de compacité acceptable, pour des valeurs faibles de *minsup*. Il est important de souligner que ce type de bases est considéré comme "difficile", vu que les approches basées sur les fermetures n'apportent pas de gains intéressants sur telles bases. Il est aussi à souligner que la taille de la représentation concise basée sur les itemsets essentiels fréquents dépasse même celle de l'ensemble de tous les itemsets fréquents. En effet, les éléments de $\mathcal{IEF}_{\mathcal{K}}$ sont caractérisés uniquement par leurs supports disjonctifs alors que ceux de BD^+ le sont uniquement par leurs supports conjonctifs. Pour cette base, quelques éléments font alors partie des deux ensembles et jouent à chaque fois un rôle différent suivant leurs appartenances.
5. Finalement, il est nettement remarquable que la variation de la cardinalité de $\mathcal{IFD}_{\mathcal{K}}$ est moins sensible à la variation de *minsup* que les autres représentations concises.

8 Conclusion et perspectives

Dans ce papier, nous avons introduit une nouvelle représentation concise exacte des itemsets fréquents. Ceci a nécessité la mise en place d'un opérateur de fermeture disjonctive permettant le calcul des éléments de la représentation, à savoir les itemsets essentiels fermés. Nous avons aussi présenté un algorithme, appelé D-CLOSURE, dédié à l'extraction de cette représentation. Les expérimentations que nous avons menées ont permis de mettre en exergue la compacité de notre représentation.

[3] Ces bases sont disponibles à partir de l'adresse : *http ://fimi.cs.helsinki.fi/data.*

minsup (%)	$\mid \mathcal{IF}_{\mathcal{K}} \mid$	$\mid \mathcal{IFF}_{\mathcal{K}} \mid$	$\mid \mathcal{IEF}_{\mathcal{K}} \mid + \mid BD^{+} \mid$	$\mid \mathcal{IFD}_{\mathcal{K}} \mid$
90	27 127	3 486	176 + 222 = 398	**22**
80	533 975	15 107	304 + 673 = 977	**83**
70	4 129 839	35 875	490 + 1 220 = 1 710	**161**
60	21 250 671	68 343	822 + 2 103 = 2 925	**265**
50	88 324 400	130 122	1 316 + 3 748 = 5 064	**462**
40	339 915 255	239 372	1 948 + 6 213 = 8 161	**819**
30	1 331 673 367	460 356	3 044 + 11 039 = 14 083	**1 625**
20	2 751 821 771	758 483	6 620 + 18 395 = 25 015	**4 725**

TAB. 4 – *La taille des différentes représentations concises dans* CONNECT.

minsup (%)	$\mid \mathcal{IF}_{\mathcal{K}} \mid$	$\mid \mathcal{IFF}_{\mathcal{K}} \mid$	$\mid \mathcal{IEF}_{\mathcal{K}} \mid + \mid BD^{+} \mid$	$\mid \mathcal{IFD}_{\mathcal{K}} \mid$
40	565	140	110 + 41 = 151	**79**
30	2 735	427	247 + 63 = 310	**182**
20	53 583	1 197	1 100 + 158 = 1 258	**759**
10	574 431	4 885	5 983 + 547 = 6 530	**4 432**
5	3 755 511	12 843	22 965 + 1 442 = 24 407	**17 814**
1	90 751 401	51 640	230 474 + 6 768 = 237 242	**186 273**

TAB. 5 – *La taille des différentes représentations concises dans* MUSHROOM.

minsup (%)	$\mid \mathcal{IF}_{\mathcal{K}} \mid$	$\mid \mathcal{IFF}_{\mathcal{K}} \mid$	$\mid \mathcal{IEF}_{\mathcal{K}} \mid + \mid BD^{+} \mid$	$\mid \mathcal{IFD}_{\mathcal{K}} \mid$
90	622	498	84 + 34 = 118	**40**
80	8 227	5 083	241 + 226 = 467	**129**
70	48 731	23 892	591 + 891 = 1 482	**349**
60	254 944	98 392	1 314 + 3 322 = 4 636	**773**
50	1 272 932	369 450	2 809 + 11 463 = 14 272	**1 685**
40	6 439 702	1 361 157	5 977 + 38 050 = 44 027	**3 563**
30	37 282 962	5 316 467	13 153 + 134 624 = 147 777	**7 957**

TAB. 6 – *La taille des différentes représentations concises dans* CHESS.

minsup (%)	$\mid \mathcal{IF}_{\mathcal{K}} \mid$	$\mid \mathcal{IFF}_{\mathcal{K}} \mid$	$\mid \mathcal{IEF}_{\mathcal{K}} \mid + \mid BD^{+} \mid$	$\mid \mathcal{IFD}_{\mathcal{K}} \mid$
5	10	10	10 + 10 = 20	**10**
4	26	26	26 + 26 = 52	**26**
3	60	60	60 + 60 = 120	**60**
2	155	155	155 + 155 = 310	**155**
1	385	385	385 + 370 = 755	**385**
0.75	561	561	561 + 453 = 1 014	**561**
0.5	1 073	1 073	1 073 + 585 = 1685	**1 073**
0.4	2 001	1 993	2 001 + 761 = 2 761	**2 001**
0.3	4 552	4 510	4 474 + 1 293 = 5 767	**4 474**
0.2	13 255	13 107	12 949 + 1 939 = 14 888	**12 944**
0.1	27 532	26 806	26 687 + 4 054 = 30 714	**26 667**

TAB. 7 – *La taille des différentes représentations concises dans* T10I4D100K.

La principale caractéristique de cette représentation est qu'elle permet d'inspirer une piste intéressante que nous souhaitons explorer. Cette piste a pour but de combler le vide entre les travaux sur les représentations concises, d'une part, et ceux sur les règles d'association génériques, d'autre part. En effet, la représentation concise introduite permet de dériver aisément les supports disjonctifs et négatifs des itemsets. Ainsi, la génération des règles (génériques) d'association généralisées deviendrait quasi-immédiate.

Remerciements

Les auteurs tiennent à remercier les relecteurs anonymes pour leurs remarques constructives. Ce travail est soutenu par le projet Franco-Tunisien CMCU 05G1412.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, et C. Zaniolo (Eds.), *Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile*, pp. 478–499.

Calders, T. et B. Goethals (2002). Mining all non-derivable frequent itemsets. In T. Elomaa, H. Mannila, et H. Toivonen (Eds.), *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2002), Springer-Verlag, LNCS, volume 2431, Helsinki, Finland*, pp. 74–85.

Calders, T., C. Rigotti, et J.-F. Boulicaut (2006). A survey on condensed representations for frequent sets. In *Constraint Based Mining, Springer-Verlag, LNAI, volume 3848*, pp. 64–80.

Casali, A., R. Cicchetti, L. Lakhal, et S. Lopes (2005). Couvertures parfaites des motifs fréquents. *Numéro thématique "Base de données avancées pour XML et le web" de la revue Ingénierie des Systèmes d'Information (ISI) 10*(2), 117–138.

Galambos, J. et I. Simonelli (2000). *Bonferroni-type inequalities with applications*. Springer.

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

Hamrouni, T., S. Ben Yahia, et E. Mephu Nguifo (2006). A new exact concise representation based on disjunctive closure. *To appear in Proceedings of the 2nd Jordan International Conference on Computer Science and Engineering (JICCSE 2006), Al-Balqa, Jordan*.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Efficient mining of association rules using closed itemset lattices. *Journal of Information Systems 24*(1), 25–46.

Summary

The interest in a further pruning of the set of frequent itemsets that can be drawn from real-life datasets is growing up. This fact is witnessed by the proliferation of what is called concise representations, *e.g.*, closed itemsets, non-derivable itemsets and essential itemsets. In this paper, we introduce a new exact concise representation that permits to drastically reduce the number of handled itemsets. Standing at the crossroads of closed itemsets and essential ones, the introduced concise representation required the definition of a new closure operator. A new algorithm, called D-CLOSURE, allowing the extraction of our representation is also presented. Carried out experiments showed an important losslessly reduction of the extracted itemsets *vs* those performed by both concise representations based on closed and essential itemsets, respectively.