Génération de bases de transactions synthétiques : vers la prise en compte des bordures

Didier Devaurs Fabien De Marchi
LIRIS, UMR CNRS 5205,
Université Lyon 1, bâtiment Nautibus
8 boulevard Niels Bohr, 69622 Villeurbanne, France

Résumé

De très divers algorithmes sont dédiés à la découverte de motifs fréquents dans les bases de données de transactions. Des initiatives collectives visant à effectuer des comparaisons de performances rigoureuses et impartiales ont vu récemment le jour. Curieusement, cette tâche est rendue difficile par le manque de jeux d'essais publics disponibles, et d'outils pour en synthétiser de façon pertinente. En particulier, un paramètre crucial conditionnant le déroulement de nombreux algorithmes est la distribution des bordures des motifs fréquents. Une seule proposition, à notre connaissance, a récemment effectué un pas vers la génération de jeux d'essais prenant en compte ce paramètre.
Dans cet article, nous étudions de près les bordures générées par la proposition existante. Une amélioration est apportée dans les calculs effectués, permettant de réduire la complexité de la génération des bases. Bien que la distribution de la bordure positive en entrée soit parfaitement respectée, nous donnons un résultat attestant que la bordure négative correspondante est toujours du même type, très différente des bordure négatives dans les bases réelles existantes. Nous esquissons alors une méthode de génération de bases synthétiques en fonction d'une distribution de bordure négative.

Mots-clés : Fouille de données, découverte des motifs fréquents, bordures, génération de bancs d'essais.

1 Introduction

La découverte des motifs fréquents dans une base de données de transaction est une des tâches de fouille de données les plus étudiées [12] ; les motifs fréquents possèdent de multiples applications [18], dont la plus connue est un élagage efficace de l'espace de construction des règles d'associations [1]. De nombreux algorithmes ont été proposés pour résoudre cette tâche, dont [2, 3, 13, 20]. Plusieurs algorithmes ont également été proposés pour la découverte des motifs fréquents maximaux par inclusion, qui sont une représentation condensée des motifs fréquents (avec toutefois perte de la valeur du support) [4, 15, 6, 11].

Devant la multitude de propositions, chaque nouvel algorithme se doit d'être soigneusement comparé à l'existant par ses auteurs ; la plupart des implémentations sont d'ailleurs disponibles sur le site web de FIMI (Frequent Itemsets Mining Implementation) [9], dont le but est de proposer une comparaison rigoureuse des propositions existantes. Les résultats soulignent que, malgré la domination moyenne de quelques outils, on ne peut pas réellement élire de "meilleur" algorithme à coup sûr ; comme prévu, certains se montrent plus ou moins adaptés à certains jeux

de données.

Toutefois, l'évaluation et la classification des algorithmes est soumise à l'existence de bases de tests nombreuses et surtout diverses dans leurs caractéristiques. Le constat est aujourd'hui fait que la communauté scientifique ne dispose pas de bancs d'essais complets et pertinents pour ce problème. Les bases utilisées par tous se limitent à une poignée de bases réelles (libres de droits), et quelques bases synthétiques, qui ont été également répertoriées sur le site web de FIMI. L'accès à plus de bases de données réelles est soumis aux contraintes des droits de propriétés et de confidentialités, et quoi qu'il en soit ne permet pas une évaluation des performances et des limites de chaque algorithme.

Cet article considère le problème de la génération de bases de données de transactions synthétiques pour l'évaluation des performances des algorithmes de découverte des motifs fréquents, maximaux ou pas. A notre connaissance, seulement peu d'attention a été accordée à ce problème jusqu'à présent. Dans [10], les principaux critères répertoriés sont le nombre de transactions, le nombre d'articles, ainsi que la "densité" (taille moyenne des transactions) des bases de données. Le nombre de transactions influe directement sur le coût des accès aux données : on observe en général des comportements linéaires en fonction de ce paramètre, avec un décrochage important lorsque la base ne tient plus en mémoire principale. Le nombre d'articles influe de façon exponentielle sur le nombre de motifs exprimables, c'est à dire la taille de l'espace de recherche, qui est exactement le treillis des parties des articles. Toutefois, le propre des algorithmes existant étant de réaliser des élagages efficaces de cet espace de recherche, l'espace réellement parcouru est en fait *indépendant du nombre d'articles*, mais bien de la disposition de l'ensemble des motifs fréquents à découvrir. C'est pourquoi la densité des jeux de données a été introduite ; mais si ce paramètre influence effectivement la taille de l'espace de recherche, il se révèle trop grossier pour le décrire finement.

Un nouveau critère plus précis a été récemment considéré dans [16]. Les auteurs étudient la notion de *distribution* des motifs fréquents, c'est à dire le nombre de motifs fréquents en fonction de leur taille. En outre, prenant en entrée une distribution quelconque pour la bordure positive, ils donnent une méthode de génération fournissant en sortie une base et un support. La bordure positive des motifs fréquents dans cette base et pour ce support possède exactement la distribution donnée en entrée. Toutefois, les auteurs ne mentionnent rien concernant la *bordure négative* des motifs fréquents ; or, nous pensons que la distribution conjointe des deux bordures est une caractéristique déterminante des jeux de données ; les observations récentes menées dans [7] tendent à confirmer cette hypothèse. Très récemment, les mêmes auteurs ont proposé une extension à leur méthode, en prenant en compte non pas une distribution de bordure positive, mais la bordure positive elle-même [17]. Dans ce cas, l'espace de recherche est alors exactement caractérisé. L'objectif annoncé des auteurs de ce travail est de savoir reproduire, de façon synthétique, les bases de données réelles existantes, pour résoudre notamment les problèmes de confidentialité. Notre objectif est différent, puisque nous cherchons à générer automatiquement des bancs d'essais très variés pour évaluer les capacités et limites des algorithmes existants.

Contribution Partant du travail de [16], et de la nature de la bordure positive générée par leur méthode, nous donnons un théorème assurant que la distribution relative des bordures positives et négatives est toujours la même. Or, nous montrons que cette distribution ne reflète pas la grande majorité des cas observés sur les jeux d'essais réels disponibles. Une méthode permettant de construire une base synthétique respectant une distribution de bordure négative donnée est alors présentée de façon préliminaire. Nous apportons également une amélioration à la mé-

thode proposée dans [16] visant à réduire la complexité des calculs effectués.

La suite de cet article est organisée de la façon suivante. La section 2 donne quelques préliminaires nécessaires à la compréhension de l'article. La section 3 synthétise une partie des travaux proposés dans [16] pour la génération de bases synthétiques, en fonction d'une distribution de bordure positive. Nous donnons dans la partie 4 un résultat visant à simplifier certains calculs dans [16]. Dans la partie 5, un théorème caractérise exactement la bordure négative des bases générées avec la méthode de [16] ; des comparaisons avec des bases réels soulignent alors les limites de cette méthode, et une méthode de génération à partir de la bordure négative est esquissée. Les conclusions et perspectives issues de ce travail sont présentées dans la section 6.

2 Préliminaires

Quelques définitions et résultats sur le thème de la découverte des motifs fréquents dans les bases de données de transactions sont introduits dans cette section. Pour plus de détails, le lecteur peut se rapporter à [12]. Les notions de bordures des motifs fréquents [19] sont également rappelées, ainsi que l'ordre Colex [5] sur des ensembles.

Bases de données de transactions, motifs fréquents et bordures Soit $\mathcal{A}$ un ensemble fini d'*articles*. Une *transaction* t est un ensemble d'articles ; une *base de données de transactions* est un multi-ensemble de transactions.

Un *motif* est un élément de $P(\mathcal{A})$, l'ensemble des parties de $\mathcal{A}$. Un motif de taille k est appelé un $k - motif$. Soit $\mathbf{d}$ une base de transactions et X un motif, le support de X dans $\mathbf{d}$ est le nombre d'occurrences de X dans $\mathbf{d}$:

$$sup(X) = |\{t \in \mathbf{d} \mid X \subseteq \mathbf{d}\}|$$

Étant donné un entier *minsup*, un motif X est dit *fréquent* si $sup(X) \geq minsup$. Soit $\mathcal{F}$ l'ensemble des motifs fréquents dans $\mathbf{d}$. Cet ensemble est clos pour la relation d'inclusion, soit

$$\forall X \in \mathcal{F}, Y \subseteq X \Longrightarrow Y \in \mathcal{F}$$

Cette caractéristique fondamentale découle de la décroissance du support des motifs par rapport à la relation d'inclusion. Ainsi, l'ensemble des motifs fréquents peut-être représenté par sa *bordure positive*[1], qui est l'ensemble de ses éléments maximaux par inclusion :

$$\mathcal{B}d^{+}(\mathcal{F}) = \{X \in \mathcal{F} \mid \forall Y \in \mathcal{F}, X \subseteq Y \Longrightarrow X = Y\}$$

D'une façon duale, $\mathcal{F}$ peut être représenté par sa *bordure négative* qui est l'ensemble des motifs non fréquents minimaux par inclusion :

$$\mathcal{B}d^{-}(\mathcal{F}) = \{X \notin \mathcal{F} \mid \forall Y \notin \mathcal{F}, Y \subseteq X \Longrightarrow X = Y\}$$

L'union de ces deux ensembles constitue la bordure de $\mathcal{F}$ notée $\mathcal{B}d(\mathcal{F})$. Remarquons que si $X \in \mathcal{B}d(\mathcal{F})$, alors tous ses sous-ensembles sont fréquents et tous ses sur-ensembles sont non fréquents.

Ordre colex Dans la suite de l'article, nous considérons que les articles sont totalement ordonnés par l'ordre lexicographique. Les articles seront nommés par leur rang dans cet ordre : $1, 2, 3....$

L'ordre colex est une alternative à l'ordre lexicographique pour ordonner totalement des ensembles de motifs de même longueur $k > 1$. Il est défini par : $X < Y$ si et seulement si il existe deux entiers $z < k$, tels que pour tout i avec $z < i \leq k$, on a $x_i = y_i$ et $x_z < y_z$.

La compréhension de l'ordre colex est fondamentales à la clarté de la suite de l'article. C'est pourquoi nous donnons l'algorithme 1, qui détaille comment générer les n premiers k-motifs de l'ordre colex.

[1] Ces motifs sont aussi connus sous le nom de *motifs fréquents maximaux*.

Algorithm 1 Génération des n premiers k-motifs

Input: n et k deux entiers ;
Output: l'ensemble C des n premiers k-motifs dans l'ordre colex

1: **for all** $i = 1$ à k **do**
2: $v[i] = i$;
3: **end for**
4: $C = \{v\}$; // *premier motif*= 1...k
5: $t = 1$;
6: **for all** $cpt = 2$ à n **do**
7: // *construction du motif suivant*
8: **while** $t < k$ et $v[t] + 1 \geq v[t+1]$ **do**
9: $t = t+1$; // *quelle position à incrémenter ?*
10: **end while**
11: $v[t] = v[t] + 1$; // *incrémentation d'une position*
12: **for all** $i = 1$ à $t - 1$ **do**
13: $v[i] = i$; // *ré-initialisation des positions précédentes*
14: **end for**
15: $C = C \cup \{v\}$;
16: **end for**
17: **return** C.

Exemple 1 Les 10 premiers 3-motifs selon l'ordre colex sont :
123, 124, 134, 234, 125, 135, 235, 145, 245, 345.

On observe alors l'une des principales caractéristiques de cet ordre, justifiant son utilisation dans ce travail : *un k-motif possède toujours le même rang, quelque soit le nombre total d'articles.* Cette propriété n'est naturellement pas vérifiée par l'ordre lexicographique. Dans la suite, l'ordre utilisé sur les motifs sera exclusivement l'ordre colex.

3 Génération de bases de tests dans [16]

Dans la suite, soit $\mathcal{A}$ un ensemble d'articles. Les notations suivantes sont empruntées à [16]. Soit $\mathcal{F}$ un ensemble de motifs ; on note $\mathcal{F}_k$ l'ensemble des k-motifs dans $\mathcal{F}$. La *représentation séquentielle* de $\mathcal{F}$ est la distribution des motifs contenus dans $\mathcal{F}$ suivant leur longueur. Elle est donnée par :

$$\prec \mathcal{F} \succ = \prec |\mathcal{F}_1|, |\mathcal{F}_2|, \ldots, |\mathcal{F}_n| \succ$$

où n est la taille du plus grand motif dans $\mathcal{F}$.

Partant d'une représentation séquentielle S, [16] propose de créer une base de données qui, une fois traitée par un algorithme de fouille de données, conduit à l'obtention d'un ensemble de motifs fréquents maximaux ayant pour représentation séquentielle la séquence S. Cette section détaille les points importants de la méthode utilisée.

3.1 Ensemble de motifs induits

Soit $\mathcal{F}$ un ensemble de motifs, les motifs induits par $\mathcal{F}$ sont tous les sous-ensembles des éléments de $\mathcal{F}$. L'utilisation de l'ordre colex permet de caractériser exactement ces motifs, *lorsque $\mathcal{F}$ est constitué des premiers motifs consécutifs d'une taille donnée*, en utilisant la notion de coefficients binomiaux.

***Lemme 1** [14] Étant donnés deux entiers n et l, n peut être écrit de façon unique sous la forme : $n = \sum_{i=t}^{l} \binom{a_i}{i} = \binom{a_l}{l} + \binom{a_{l-1}}{l-1} + \binom{a_{l-2}}{l-2} + \cdots + \binom{a_t}{t}$ où $t \geq 1$, et $a_l > a_{l-1} > \cdots > a_t$ sont des entiers naturels tels que : $\forall\ i\ : t \leq i \leq l,\ a_i \geq i$.*

Autrement dit, tout entier n peut être écrit de façon unique comme une somme de coefficients binomiaux, appelée *l-représentation canonique de n* dans [8], que nous noterons dans la suite : $LB_l^0(n)$.
Nous noterons également : pour $1 \leq k < l$,
$LB_l^k(h) = \binom{a_l}{l-k} + \binom{a_{l-1}}{l-1-k} + \cdots + \binom{a_t}{t-k}$
où $\binom{a}{b} = 0$ si $b < 0$.

Les nombres a_i sont calculés de la façon suivante :

- l'entier a_l est tel que $\binom{a_l}{l} \leq n < \binom{a_l+1}{l}$
- a_{l-1} est tel que $\binom{a_{l-1}}{l-1} \leq n - \binom{a_l}{l} < \binom{a_{l-1}+1}{l-1}$

- ...
- a_i vérifie $\binom{a_i}{i} \leq n - \sum_{j=i+1}^{l} \binom{a_j}{j} < \binom{a_i+1}{i}$
- ...

Exemple 2 Calculons la 3-représentation canonique de 5 (i.e. $n = 5$ et $l = 3$). On a : $\binom{3}{3} = 1$, $\binom{4}{3} = 4$ et $\binom{5}{3} = 10$. D'où : $\binom{4}{3} < 5 < \binom{5}{3}$, et donc $a_3 = 4$. On a maintenant : $5 - \binom{4}{3} = 1$, et $1 = \binom{2}{2}$. D'où : $a_2 = 2$. Comme $5 - \binom{4}{3} - \binom{2}{2} = 0$, le processus s'arrête. On a donc : $t = 2$, $a_3 = 4$, $a_2 = 2$ et $LB_3^0(5) = \binom{4}{3} + \binom{2}{2}$.

Considérons maintenant deux entiers n et l. Soit $\mathcal{F}$ l'ensemble des n premiers l-motifs dans l'ordre colex. Nous allons déterminer le nombre de motifs induits par l'ensemble $\mathcal{F}$. Les résultats sont donnés par les lemmes suivants :

***Lemme 2** [5] Les $(l-1)$-motifs induits par $\mathcal{F}$ sont les $LB_l^1(n)$ premiers $(l-1)$-motifs dans l'ordre colex.*

***Lemme 3** [16] Soit k un entier tel que $1 \leq k < l$. Les k-motifs induits par $\mathcal{F}$ sont les $LB_l^{l-k}(n)$ premiers k-motifs dans l'ordre colex.*

Exemple 3 Reprenons l'exemple précédent, et considérons les 5 premiers 3-motifs dans l'ordre colex : $\mathcal{F} = \{123, 124, 134, 234, 125\}$. Nous avons vu que : $LB_3^0(5) = \binom{4}{3} + \binom{2}{2}$. D'où : $LB_3^{3-2}(5) = LB_3^1(5) = \binom{4}{2} + \binom{2}{1} = 8$. Donc, les 2-motifs induits par $\mathcal{F}$ sont les 8 premiers 2-motifs dans l'ordre colex : 12, 13, 23, 14, 24, 34, 15 et 25. De même, $LB_3^{3-1}(5) = LB_3^2(5) = \binom{4}{1} + \binom{2}{0} = 5$. Donc, les 1-motifs (i.e. les articles) induits par $\mathcal{F}$ sont les 5 premiers 1-motifs dans l'ordre colex : 1, 2, 3, 4 et 5.

Nous venons de présenter une façon de calculer le nombre de motifs induits par un ensemble de motifs $\mathcal{F}$ donné. En pratique, nous serons plutôt amenés à calculer le nombre de motifs induits par plusieurs ensembles de motifs donnés. Autrement dit, si l'on dispose de h ensembles de motifs F_1, F_2, ..., F_h (avec $h \geq 2$), on aimerait connaître l'ensemble des motifs X tels que : $\exists\ i\ : 1 \leq i \leq h$, X est un sous-ensemble d'un élément de F_i. Le lemme suivant explicite ce résultat dans le cas de deux ensembles de motifs. Le théorème qui suit est une généralisation au cas où h (avec $h \geq 2$) ensembles de motifs sont donnés.

***Lemme 4** [16] Soient A l'ensemble des n_A premiers l_A-motifs et B l'ensemble des n_B premiers l_B-motifs dans l'ordre colex. Pour tout entier k tel que : $1 \leq k < min(l_A, l_B)$, l'ensemble des k-motifs induits conjointement par A et B sont les m premiers k-motifs dans l'ordre colex, où $m = max(LB_{l_A}^{l_A-k}(n_A), LB_{l_B}^{l_B-k}(n_B))$.*

Exemple 4 Soient $A = \{123, 124, 134, 234, 125\}$ (i.e. $n_A = 5$ et $l_A = 3$), et $B = \{1234, 1235, 1245\}$ (i.e. $n_B = 3$ et $l_B = 4$). On a : $LB_4^0(3) = \binom{4}{4} + \binom{3}{3} + \binom{2}{2}$. D'où : $LB_4^{4-2}(3) = LB_4^2(3) = \binom{4}{2} + \binom{3}{1} + \binom{2}{0} = 10$. Donc, les 2-motifs induits par B sont les 10 premiers 2-motifs dans l'ordre colex : 12, 13, 23, 14, 24, 34, 15, 25, 35, et 45. Or, nous avons vu que A induisait seulement les 8 premiers 2-motifs dans l'ordre colex : 12, 13, 23, 14, 24, 34, 15 et 25. Par conséquent, les 2-motifs induits conjointement par A et B sont les 10 premiers 2-motifs dans l'ordre colex : 12, 13, 23, 14, 24, 34, 15, 25, 35 et 45.

***Théorème 1** [16] Soit h un entier ($h \geq 2$). Considérons, pour tout i tel que $1 \leq i \leq h$, l'ensemble F_i des n_i premiers l_i-motifs dans l'ordre colex. Pour tout entier k tel que : $1 \leq k < min_{i=1}^{h}\{l_i\}$, l'ensemble des k-motifs induits conjointement par F_1, F_2, ..., F_h sont les $max_{i=1}^{h}\{LB_{l_i}^{l_i-k}(n_i)\}$ premiers k-motifs dans l'ordre colex.*

3.2 Méthode de génération

Utilisant ces résulats, la méthode de génération présentée dans [16] se décompose en deux étapes :

- la génération d'un ensemble de motifs maximaux pour la bordure positive
- la création d'une base de données de transactions

3.2.1 Génération d'un ensemble de motifs maximaux

Soit une séquence $S = \prec s_1, s_2, \ldots, s_p \succ$ de p entiers positifs. La première étape consiste à générer un ensemble de motifs maximaux $\mathcal{M}$ ayant pour représentation séquentielle la séquence S. Autrement dit, on souhaite que, pour chaque niveau k (où $1 \leq k \leq p$), $\mathcal{M}_k$ contienne s_k motifs de longueur k. La seule contrainte étant, pour obtenir un ensemble de maximaux correct, qu'aucun motifs ne soient comparables deux à deux selon l'inclusion.

Le théorème suivant caractérise de façon constructive un tel ensemble $\mathcal{M}$. Deux résultats importants sont à remarquer dans l'énonciation du théorème : 1) Le nombre d'articles utilisés est minimal et 2) Toute distribution peut-être acceptée en entrée.

***Théorème 2** [16]. Soient p un entier positif non nul et $S = \prec s_1, s_2, \ldots, s_p \succ$ une séquence de p entiers positifs (avec $s_p \neq 0$). Soient $r_1, r_2, \ldots, r_p$ p entiers positifs tels que : $r_p = 0$ et $\forall\ 1 \leq k < p,\ r_k = max_{j=k+1}^{p}\{LB_j^{j-k}(r_j + s_j)\}$. Considérons $\mathcal{M}$ l'ensemble de motifs construit de la façon suivante : $\forall\ 1 \leq k \leq p$, ajoutons à $\mathcal{M}$ les s_k k-motifs de rangs $r_k+1, r_k+2, \ldots, r_k+s_k$ dans l'ordre colex. Alors, $\mathcal{M}$ est un ensemble de motifs maximaux tel que $\prec \mathcal{M} \succ = S$. De plus, le nombre d'articles utilisés est : $r_1 + s_1$. C'est le nombre minimum d'articles nécessaires à la construction d'un ensemble de motifs maximaux de séquence S.*

L'intuition de ce théorème est la suivante. $\mathcal{M}$ est construit niveau par niveau, en commençant par les motifs de longueur p. $\mathcal{M}$ doit contenir s_p motifs maximaux de longueur p : on prend les s_p premiers motifs de longueur p dans l'ordre colex. Or, ces motifs induisent $LB_p^1(s_p)$ motifs de longueur $p-1$ (on notera $r_{p-1} = LB_p^1(s_p)$), et ces motifs sont les r_{p-1} premiers $(p-1)$-motifs dans l'ordre colex. Puisque ce sont des sous-ensembles des motifs de longueur p, ils ne peuvent pas être maximaux. On utilisera donc les motifs suivants dans l'ordre colex. Ainsi, au niveau $p-1$, on ajoutera à $\mathcal{M}$ les $(p-1)$-motifs de rangs $r_{p-1}+1$, $r_{p-1}+2$, ..., $r_{p-1}+s_{p-1}$, qui sont bien maximaux. L'exemple suivant illustre cette méthode de génération.

Exemple 5 Soit la séquence $S = \prec 2,\ 3,\ 2 \succ$ (i.e. $p = 3$, $s_1 = 2$, $s_2 = 3$ et $s_3 = 2$). On a : $r_3 = 0$. $\mathcal{M}$ contient les 2 premiers 3-motifs dans l'ordre colex : 123 et 124. Puisque $s_3 = 2 = \binom{3}{3} + \binom{2}{2}$, on a : $r_2 = LB_3^1(s_3) = \binom{3}{2} + \binom{2}{1} = 5$. Les 5 premiers 2-motifs dans l'ordre colex sont 12, 13, 23, 14 et 24 (qui ne sont pas maximaux). On va ajouter à $\mathcal{M}$ les 3 2-motifs suivants dans l'ordre colex : 34, 15 et 25. Puisque $s_2+r_2 = 8 = \binom{4}{2} + \binom{2}{1}$, on a : $LB_2^1(s_2 + r_2) = \binom{4}{1} + \binom{2}{0} = 5$. De plus, $LB_3^2(s_3) = \binom{3}{1} + \binom{2}{0} = 4$. Ainsi, $r_1 = max(LB_2^1(s_2+r_2), LB_3^2(s_3)) = 5$ Les 5 premiers 1-motifs (i.e. les articles) sont 1, 2, 3, 4 et 5 (qui ne sont pas maximaux). On ajoute à $\mathcal{M}$ les 2 articles suivants : 6 et 7. Finalement, on obtient l'ensemble de motifs maximaux suivant : $\mathcal{M} = \{6,\ 7,\ 34,\ 15,\ 25,\ 123,\ 124\}$.

3.2.2 Génération d'une base de transactions

La génération de la base de transactions à partir d'une telle séquence pour la bordure positive est très simple : il suffit de générer une base avec un transaction pour chaque motif maximal, et donc considérer un support minimal $minsup =$

1. Le nombre de transactions obtenues est alors la somme de tous les éléments de la séquence prise en entrée ; ce nombre peut être augmenter en répliquant les transaction ou leurs sous-ensembles, sans altérer la bordure.

Remarque : Dans [16], les auteurs considèrent en fait un ensemble de séquences en entrée. Dans la base générée, toutes ces séquences correspondent à une bordure positive des fréquents, pour des valeurs différentes du seuil de support. Cette extension ne comportant pas de difficulté théorique particulière, nous préférons nous concentrer dans cet article sur le cas d'une seule séquence en entrée, pour des raisons de clarté.

4 Simplification du calcul des entiers r_k

Le calcul des entiers r_k qui interviennent dans le théorème 2 est assez complexe, puisqu'il nécessite la recherche d'un maximum sur toutes les étapes précédentes de la méthode de génération. Or, ce calcul s'appuie sur le résultat présenté dans le théorème 1, obtenu pour une collection d'ensembles $\mathcal{M}_1$, $\mathcal{M}_2$, ..., $\mathcal{M}_h$, n'ayant aucun lien entre eux. Lors de la génération d'un ensemble de motifs maximaux, on se place dans un cas particulier, où un ensemble $\mathcal{M}_i$ contient les motifs induits par l'ensemble $\mathcal{M}_{i+1}$.

De façon intuitive, on peut se rendre compte dans l'exemple 4, qu'à un niveau k donné (où $1 \leq k \leq p-1$), il est inutile de considérer les motifs induits par les ensembles construits à tous les niveaux précédents. Il suffit de déterminer le nombre de motifs induits par l'ensemble de motifs considéré au niveau $k+1$. En effet, l'ensemble des k-motifs induits par les $(k+1)$-motifs considérés au niveau $k+1$ contient l'ensemble des k-motifs induits par les $(k+2)$-motifs considérés au niveau $k+2$, qui lui-même contient l'ensemble des k-motifs induits par les $(k+3)$-motifs considérés au niveau $k+3$, et ainsi de suite, jusqu'au niveau p. Ceci est exprimé à l'aide de la proposition suivante :

Proposition 1 *Soient p un entier positif non nul et $S = \prec s_1, s_2, \ldots, s_p \succ$ une séquence de p entiers positifs (avec $s_p \neq 0$). Considérons p entiers positifs r_1, r_2, ..., r_p définis par : $r_p = 0$ et $\forall\ 1 \leq k \leq p-1$, $r_k = LB^1_{k+1}(r_{k+1} + s_{k+1})$. Soient $\mathcal{M}$ et $\mathcal{M}'$ les ensembles de motifs construits suivant le procédé décrit dans le théorème 2 en utilisant respectivement, cette nouvelle définition des entiers r_k, et la définition présentée dans [16]. On obtient $\mathcal{M} = \mathcal{M}'$.*

Preuve *Pour s'assurer de ceci, montrons que : $\forall\ 1 \leq k \leq p-1$, $r_k = LB^1_{k+1}(r_{k+1} + s_{k+1})$ en considérant que les p entiers r_1, ..., r_p sont définis par le procédé décrit dans le théorème 2, à savoir : $r_p = 0$ et $\forall\ 1 \leq k \leq p-1$,*
$r_k = max^p_{j=k+1}(LB^{j-k}_j(r_j + s_j))$.
Nous allons procéder par récurrence sur k.
- pour $k = p-1$:
j prend comme seule valeur p, ce qui donne bien :
$r_{p-1} = LB^1_p(r_p + s_p)$
- pour $k = p-2$:
$r_{p-2} = max(LB^1_{p-1}(r_{p-1} + s_{p-1}), LB^2_p(s_p))$.
Or, $r_{p-1} = LB^1_p(s_p)$
$= \binom{a_p}{p-1} + \binom{a_{p-1}}{p-2} + \cdots + \binom{a_{p-u+1}}{p-u}$,
si $s_p = \binom{a_p}{p} + \binom{a_{p-1}}{p-1} + \cdots + \binom{a_{p-u+1}}{p-u+1}$.
De plus, $r_{p-1} = LB^0_{p-1}(r_{p-1})$
$= \binom{b_{p-1}}{p-1} + \binom{b_{p-2}}{p-2} + \cdots + \binom{b_{p-v}}{p-v}$.
D'après l'unicité de cette décomposition en somme de coefficients binomiaux, on a : $u = v$ et $\forall\ 1 \leq i \leq u$, $a_{p-i+1} = b_{p-i}$.
Ainsi, $LB^2_p(s_p) = \binom{a_p}{p-2} + \binom{a_{p-1}}{p-3} + \cdots + \binom{a_{p-u+1}}{p-u-1}$
$= \binom{b_{p-1}}{p-2} + \binom{b_{p-2}}{p-3} + \cdots + \binom{b_{p-v}}{p-v-1}$
$= LB^1_{p-1}(r_{p-1})$.
Or, le nombre de $(p-1)$-motifs induits par r_{p-1} p-motifs est inférieur au nombre de $(p-1)$-motifs induits par $r_{p-1} + s_{p-1}$ p-motifs (car $s_{p-1} \geq 0$).
D'où : $LB^1_{p-1}(r_{p-1}) \leq LB^1_{p-1}(r_{p-1} + s_{p-1})$
et donc, $LB^2_p(s_p) \leq LB^1_{p-1}(r_{p-1} + s_{p-1})$

ce qui implique que $r_{p-2} = LB^1_{p-1}(r_{p-1} + s_{p-1})$.

- Hypothèse de récurrence : Supposons que pour un entier k fixé (où $1 \leq k \leq p-2$*), on ait :* $\forall\ k+1 \leq j \leq p-1,\ r_j = LB^1_{j+1}(r_{j+1} + s_{j+1})$.

- Montrons que : $r_k = LB^1_{k+1}(r_{k+1} + s_{k+1})$.

On a : $r_k = max^p_{j=k+1}(LB^{j-k}_j(r_j + s_j))$.
Soit j un entier tel que : $k+1 \leq j \leq p-1$.
On a : $LB^1_{j+1}(r_{j+1} + s_{j+1}) = r_j = LB^0_j(r_j)$,
d'où : $LB^2_{j+1}(r_{j+1} + s_{j+1}) = LB^1_j(r_j)$, ...,
et donc, $LB^{j-k+1}_{j+1}(r_{j+1} + s_{j+1}) = LB^{j-k}_j(r_j)$
De plus, $LB^{j-k}_j(r_j) \leq LB^{j-k}_j(r_j + s_j)$
D'où : $\forall\ k+1 \leq j \leq p-1$,
$LB^{j+1-k}_{j+1}(r_{j+1} + s_{j+1}) \leq LB^{j-k}_j(r_j + s_j)$.
Ainsi,
$$\begin{aligned} LB^{p-k}_p(s_p) &\leq LB^{p-1-k}_{p-1}(r_{p-1} + s_{p-1}) \\ &\leq \dots \\ &\leq LB^1_{k+1}(r_{k+1} + s_{k+1}). \end{aligned}$$
Donc, $r_k = LB^1_{k+1}(r_{k+1} + s_{k+1})$.

- Conclusion : Par récurrence, on obtient : $\forall\ 1 \leq k \leq p-1,\ r_k = LB^1_{k+1}(r_{k+1} + s_{k+1})$ □

Exemple 6 Soit la séquence $S = \prec 2,\ 3,\ 2 \succ$ (i.e. $p = 3$, $s_1 = 2$, $s_2 = 3$ et $s_3 = 2$). On a : $r_3 = 0$. Les 2 premiers 3-motifs dans l'ordre colex sont 123 et 124. Puisque $s_3 = 2 = \binom{3}{3} + \binom{2}{2}$, on a : $r_2 = LB^1_3(s_3) = \binom{3}{2} + \binom{2}{1} = 5$. Les 5 premiers 2-motifs dans l'ordre colex sont 12, 13, 23, 14 et 24. Les 3 suivants sont 34, 15 et 25. Puisque $s_2 + r_2 = 8 = \binom{4}{2} + \binom{2}{1}$, on a : $r_1 = LB^1_2(s_2 + r_2) = \binom{4}{1} + \binom{2}{0} = 5$. Les 5 premiers 1-motifs sont 1, 2, 3, 4 et 5, et les 2 suivants sont 6 et 7. Ces résultats peuvent être représentés à l'aide du treillis donné en figure 4.

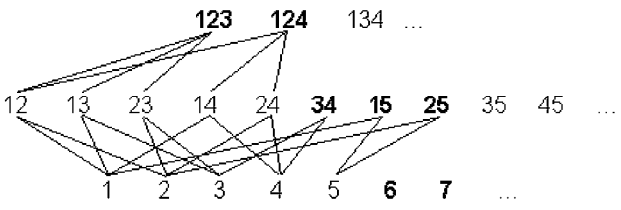

FIG. 1 – Illustration du calcul des motifs induits

5 Prise en compte de la bordure négative

En partant d'une séquence S d'entiers positifs donnés, la méthode présentée dans [16] permet de générer une bordure positive de fréquents ayant pour représentation séquentielle la séquence S. Or, un autre paramètre important d'un jeux de données est la distribution de la bordure négative des fréquents ; en effet, ce sont en général les premiers (et souvent les seuls) motifs non fréquents parcourus par la quasi-totalité des algorithmes de fouille. Nous caractérisons ici la distribution de la bordure négative correspondante générée, et comparons alors ce résultat aux bases de données réelles existantes. Nous donnons ensuite un moyen de prendre en entrée une distribution de bordure négative.

5.1 Calcul de la bordure négative synthétique

Comment peut-on, à partir de $\prec \mathcal{B}d^+(\mathcal{F}) \succ$ et de la méthode de génération présentée plus haut, en déduire $\prec \mathcal{B}d^-(\mathcal{F}) \succ$? Avant d'énoncer le théorème établissant cette correspondance, nous en donnons l'intuition pour les premiers niveaux.

Notons $\prec \mathcal{B}d^-(\mathcal{F}) \succ = \prec t_1,\ t_2,\ \cdots \succ$. Puisque tous les articles utilisés sont dans F, on a : $t_1 = 0$ (i.e. on n'introduit aucun article non fréquent).

Le nombre d'articles fréquents est $r_1 + s_1$: ce sont les entiers 1, 2, ..., $r_1 + s_1$. On sait que les entiers 1, 2, ..., r_1 sont induits par les $r_2 + s_2$ premiers 2-motifs dans l'ordre colex. Le but est maintenant de déterminer l'ensemble des 2-motifs induisant les entiers 1, 2, ..., $r_1 + s_1$. Il s'agit tout simplement de leurs sur-ensembles de taille 2, qui sont au nombre de $\binom{r_1+s_1}{2}$. De plus, on sait que ces 2-motifs sont soit des motifs fréquents, soit des éléments de la bordure négative (puisque tous leurs sous-ensembles sont fré-

quents), que ce sont les premiers 2-motifs dans l'ordre colex (car ils induisent les $r_1 + s_1$ premiers articles dans l'ordre colex), et qu'ils sont au nombre de $r_2+s_2+t_2$. On a donc $r_2+s_2+t_2 = \binom{r_1+s_1}{2}$. Par conséquent, $t_2 = \binom{r_1+s_1}{2} - (r_2 + s_2)$.

Notation. Étant donnés deux entiers n et l, si la l-représentation canonique de n est $LB_l^0(n) = \binom{a_l}{l}+\binom{a_{l-1}}{l-1}+\cdots+\binom{a_t}{t}$, nous noterons $LB_l^{-1}(n) = \binom{a_l}{l+1}+\binom{a_{l-1}}{l}+\cdots+\binom{a_t}{t+1}$, où $\binom{a}{b} = 0$, si $b > a$.

Théorème 3 *Soient p un entier positif non nul et $S = \prec s_1, s_2, \ldots, s_p \succ$ une séquence de p entiers positifs (avec $s_p \neq 0$). Soient $r_1, r_2, \ldots, r_p$ p entiers positifs tels que : $r_p = 0$ et $\forall\ 1 \leq k \leq p-1,\ r_k = LB_{k+1}^1(r_{k+1} + s_{k+1})$. Considérons $\mathcal{M}$, l'ensemble de motifs maximaux construit suivant le procédé décrit dans le théorème 1, et $\mathcal{F}$ l'ensemble de motifs ayant $\mathcal{M}$ pour bordure positive (i.e. tel que $\prec \mathcal{B}d^+(\mathcal{F}) \succ = S$). Si l'on considère les $p+1$ entiers $t_1, t_2, \ldots, t_p, t_{p+1}$ définis par : $t_1 = 0$ et $\forall\ 1 \leq k \leq p,\ t_{k+1} = LB_k^{-1}(r_k + s_k) - (r_{k+1} + s_{k+1})$ (en posant $s_{p+1} = r_{p+1} = 0$), on a :*
$\prec \mathcal{B}d^-(\mathcal{F}) \succ = \prec t_1, t_2, \ldots, t_p, t_{p+1} \succ$.

Preuve *Puisque tous les articles utilisés sont induits par les éléments de la bordure positive, la bordure négative ne contient aucun motif de taille 1.*
Soit k un entier tel que : $1 \leq k \leq p$. On sait que les $r_k + s_k$ premiers k-motifs dans l'ordre colex sont dans $\mathcal{F}$. Leurs sur-ensembles de taille $k+1$ sont donc soit dans $\mathcal{F}$, soit dans $\mathcal{B}d^-(\mathcal{F})$. On sait que les $r_{k+1} + s_{k+1}$ premiers $(k+1)$-motifs dans l'ordre colex sont dans $\mathcal{F}$. Considérons maintenant les t_{k+1} suivants, où $t_{k+1} = LB_k^{-1}(r_k + s_k) - (r_{k+1} + s_{k+1})$. On a alors : $r_{k+1} + s_{k+1} + t_{k+1} = LB_k^{-1}(r_k + s_k)$. D'après l'unicité de la $(k+1)$-représentation canonique de $r_{k+1} + s_{k+1} + t_{k+1}$, la décomposition en coefficients binomiaux donnée par $LB_k^{-1}(r_k + s_k)$ est exactement la même que celle donnée par $LB_{k+1}^0(r_{k+1} + s_{k+1} + t_{k+1})$. Donc, les k-motifs induits par l'ensemble des $r_{k+1}+s_{k+1}+t_{k+1}$ premiers $(k+1)$-motifs dans l'ordre colex sont les $LB_{k+1}^1(r_{k+1}+s_{k+1}+t_{k+1}) = LB_k^0(r_k+s_k) = r_k+s_k$ premiers k-motifs dans l'ordre colex. Ainsi, les $(k+1)$-motifs ayant pour rang $r_{k+1}+s_{k+1}+1$, $r_{k+1} + s_{k+1} + 2, \ldots, r_{k+1} + s_{k+1} + t_{k+1}$ dans l'ordre colex, sont dans $\mathcal{B}d^-(\mathcal{F})$. De plus, ce sont les seuls, puisque les $(k+1)$-motifs de rangs 1, 2, $\ldots$, $r_{k+1} + s_{k+1}$ sont dans $\mathcal{F}$, et que ceux de rang supérieur à $r_{k+1}+s_{k+1}+t_{k+1}$ n'ont pas tous leurs sous-ensembles dans $\mathcal{F}$.

□

Exemple 7 Soit la séquence $S = \prec 2,\ 3,\ 2 \succ$. On a : $\mathcal{M} = \{6,\ 7,\ 34,\ 15,\ 25,\ 123,\ 124\}$ et $\mathcal{F} = \{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 12,\ 13,\ 23,\ 14,\ 24,\ 34,\ 15,\ 25,\ 123,\ 124\}$, c'est-à-dire que $r_3 = 0$, $r_2 = 5$ et $r_1 = 5$. Le théorème 3 nous donne : $t_1 = 0$. On a : $r_1+s_1 = 7 = \binom{7}{1}$. D'où : $t_2 = \binom{7}{2} - (r_2+s_2) = 21 - 8 = 13$. De même, $r_2 + s_2 = 8 = \binom{4}{2} + \binom{2}{1}$. Donc, $t_3 = \binom{4}{3} + \binom{2}{2} - (r_3 + s_3) = 4 + 1 - 2 = 3$. Enfin, $r_3 + s_3 = 2 < \binom{4}{3}$, et donc $t_4 = 0$. En partant de $\prec \mathcal{B}d^+(\mathcal{F}) \succ = \prec 2,\ 3,\ 2 \succ$, on obtient donc $\prec \mathcal{B}d^-(\mathcal{F}) \succ = \prec 0,\ 13,\ 3 \succ$ et $\mathcal{B}d^-(\mathcal{F}) = \{35,\ 45,\ 16,\ 26,\ 36,\ 46,\ 56,\ 17,\ 27,\ 37,\ 47,\ 57,\ 67,\ 134,\ 234\}$. Ces résultats peuvent être représentés à l'aide du treillis donné en figure 2. Notez que, dans cette figure, tous les motifs non représentés sont non fréquents.

5.2 Comparaison avec les bases de données existantes

Nous avons implantés le calcul de la distribution de la bordure négative synthétique à partir d'une distribution de bordure positive, selon le théorème 3. Les tests réalisés sont alors construits de la façon suivante.

En entrée, nous considérons les distributions des bordures positives de certaines bases de données classiquement utilisées et disponibles sur le site de FIMI, pour des supports donnés.

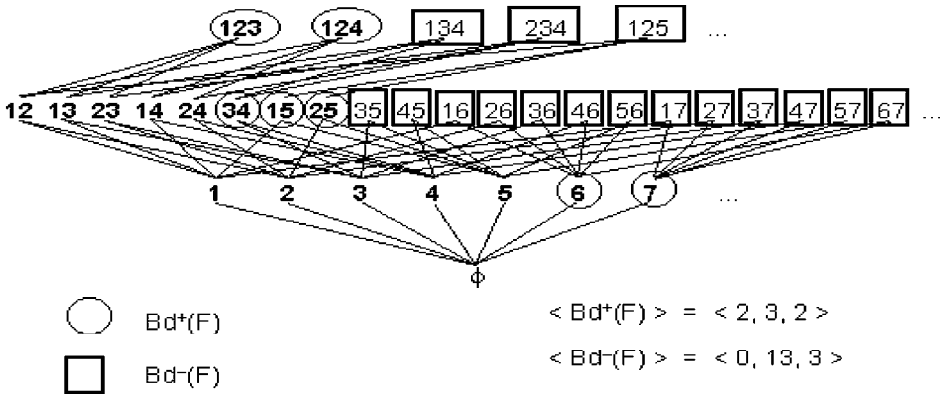

FIG. 2 – Illustration du calcul de la bordure négative

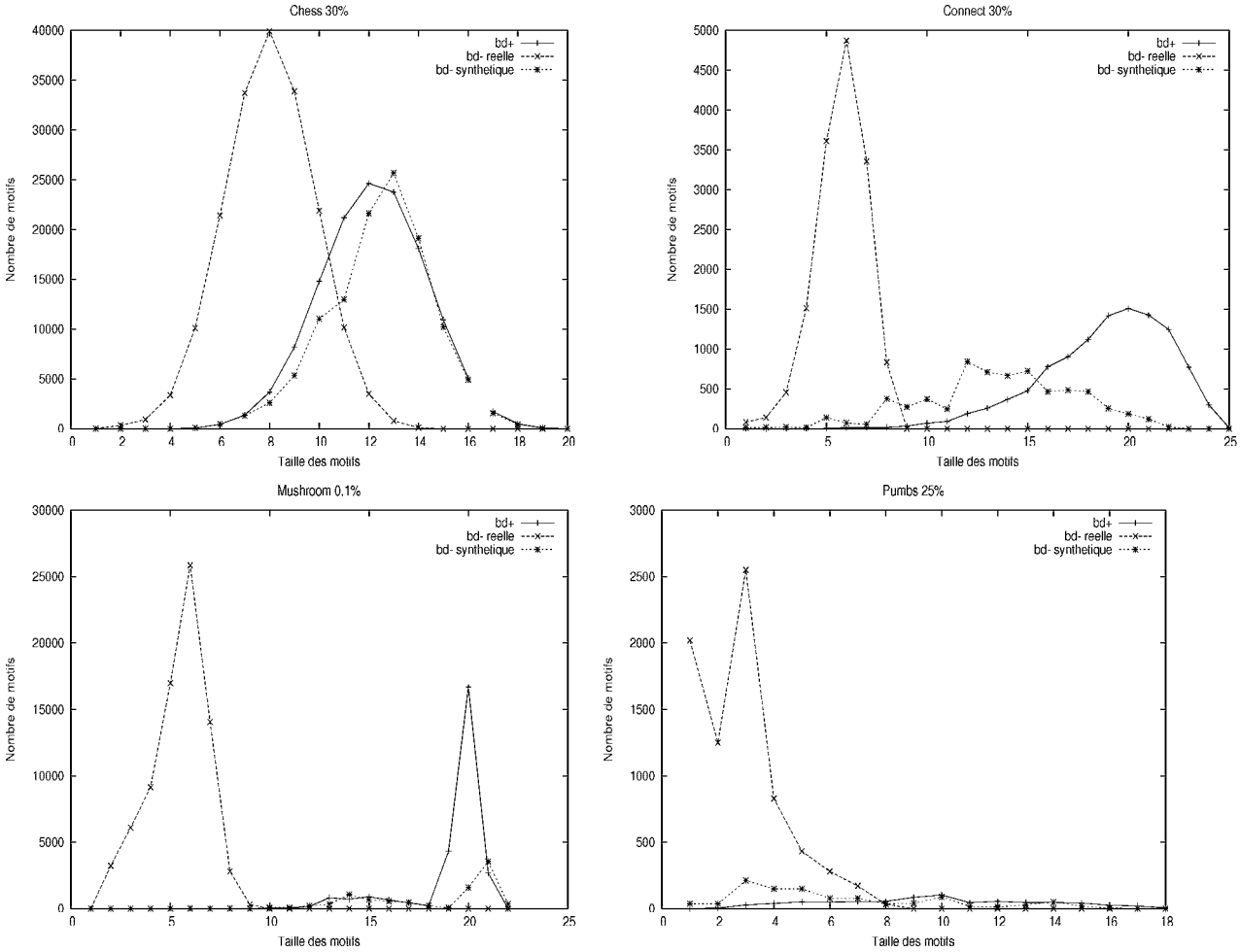

FIG. 3 – Comparaison entre les bordures négatives réelles et synthétiques, pour des bordures positives réelles

En utilisant [16], on peut construire des bases synthétiques possédant exactement la même distribution de bordure positive des fréquents ; ce protocole est exactement celui utilisé dans [16] pour la validation. A partir de cette bordure positive, nous calculons la distribution de la bordure négative correspondante pour les bases synthétiques (théorème 2). Cette bordure synthétique peut alors être comparée à la bordure négative réelle découverte dans les bases originales. Les résultats sont reportés dans la figure 3.

On observe que pour l'ensemble des bases de données considérées, la distribution de la bordure négative synthétique est radicalement différente de la bordure négative réelle. Dans tous les cas, la bordure négative synthétique est beaucoup plus proche de la bordure positive que dans la normale. Or, dans certains cas, la bordure négative réelle est exclusivement concentrée dans les premiers niveaux ; cette caractéristique est d'ailleurs assez typique d'un certain nombre de jeux de données existants [7]. Il est évident que la nature de la bordure négative est un paramètre majeur pour l'efficacité des algorithmes existants ; en effet, la plupart utilisent un élagage basé sur la découverte d'éléments de cette bordure. Les complexités exhibées par exemple dans [19, 11] pour la découverte des motifs fréquents et des motifs maximaux fréquents sont exprimées notamment en fonction du nombre d'éléments dans la bordure négative.

Ces résultats mettent donc en évidence que le problème de la génération de bases synthétiques ne peut s'arrêter à la seule prise en compte de la bordure positive. Un travail doit maintenant être mené pour permettre des tests selon différentes bordures négatives, ainsi que selon différentes positions relatives des deux bordures.

5.3 Une bordure négative en entrée

Nous esquissons ici, de façon informelle, une méthode de génération, considérant en entrée une distribution pour la bordure négative.

Remarquons pour commencer que *toute distribution est réalisable.* En effet, en utilisant la méthode du théorème 2, on peut toujours générer un ensemble de motifs non comparables deux à deux par inclusion, constituant donc une bordure négative potentielle d'un ensemble de motifs fréquents. En outre, rappelons que ce même théorème 2 utilise un nombre minimal d'articles, propriété qu'il est souhaitable de conserver pour donner plus de souplesse au générateur.

Nous illustrons dans la suite, au travers d'un exemple, la méthode que nous prévoyons pour générer une base de transactions synthétique à partir d'une séquence en entrée. Il s'agit d'un travail préliminaire ; une formalisation sous la forme de théorème, et une preuve de correction sont en cours.

Soit $T = \prec t_1, ..., t_p \succ$ une séquence de p entiers en entrée. La méthode se décompose en trois étapes détaillées ci-dessous.

Détermination du nombre minimal d'articles nécessaires Pour cela, la première étape applique le théorème 2 à la séquence T. On obtient donc en sortie, notamment, le nombre minimal d'articles nécessaires pour construire un ensemble de motifs de séquence T incomparables deux à deux par inclusion, que l'on notera n_1. L'exemple suivant, accompagné de la figure 4, illustrent cette étape.

Exemple 8 Soit la séquence $T = \prec 0, 2, 3, 1 \succ$. Puisque $t_4 = 1$, on considère le premier 4-motif dans l'ordre colex : 1234. Il induit les $LB_4^1(t_4) =$ 4 premiers 3-motifs dans l'ordre colex : 123, 124, 134 et 234. A ces motifs on ajoute les trois suivants, puisque $t_3 = 3$, soit 125, 135 et 235.
Ces sept 3-motifs induisent les $LB_3^1(7) = 9$ premiers 2-motifs dans l'ordre colex : 12, 13, 23, 14, 24, 34, 15, 25 et 35, auxquels on ajoute les 2 motifs suivants ($t_2 = 2$) : 45 et 16.
L'ensemble de ces 2-motifs induit les $LB_2^1(11) =$ 6 premiers 1-motifs dans l'ordre colex : 1, 2, 3,

4, 5 et 6. Puisque $t_1 = 0$, on ne rajoute pas d'articles.
Le nombre d'articles nécessaires est donc $n_1 = 6$.

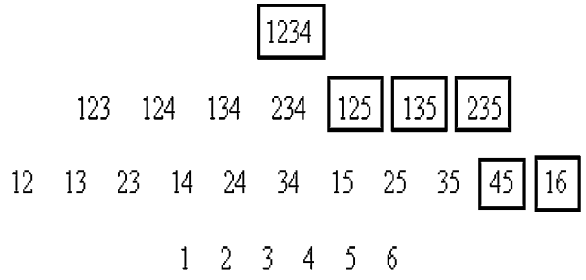

FIG. 4 – Etape 1 : Détermination du nombre d'articles

Pourquoi ne pas s'arrêter, et générer une base possédant cette bordure négative ? Le problème serait alors de déterminer la bordure positive correspondante. En effet, la connaissance de la bordure positive est nécessaire pour la construction de la base de tests. Or, dans la figure 4, on peut observer par exemple que le motif 26 n'a pas été considéré alors que tous ses sous-ensembles sont fréquents : c'est un élément qui doit-être dans la bordure positive. Ce cas de figure, généralisable à tous les niveaux, rendrait très complexe la génération de la base.

Détermination des motifs de la bordure négative Au premier niveau, on insère dans $\mathcal{B}d^-(\mathcal{F})$ les t_1 motifs de l'intervalle $[n_1 - t_1 + 1; n_1]$. Les autres sont considérés comme fréquents dans la base de transactions à construire.

Puis à chaque niveau $k \geq 2$ on considère les n_k premiers k-motifs comme étant l'ensemble des motifs dont tous les sous-motifs de taille $k-1$ sont fréquents, donc leur rang est dans $[1; n_{k-1} - t_{k-1}]$. D'après ce qui précède, $n_k = LB_{k-1}^{-1}(n_{k-1} - t_{k-1})$. On prend alors les motifs de rang $[n_k - t_k + 1; n_k]$ et on les insère dans $\mathcal{B}d^-(\mathcal{F})$. Les motifs restants parmi les n_k sont considérés comme fréquents dans la future base de transactions. Enfin, les motifs supérieurs à n_k sont considérés non fréquents, ainsi que tous les motifs des niveaux non considérés.

Ainsi, on peut vérifier facilement que tous les sous-motifs des élément insérés dans $\mathcal{B}d^-(\mathcal{F})$ sont fréquents, alors que tous leurs sur-motifs sont non fréquents. Si l'on suppose ces éléments non fréquents, ils constituent donc bien la bordure négative. L'exemple suivant et la figure 5 illustrent cette étape.

Exemple 9 Poursuivons l'exemple précédent. On a : $n_1 = 6$. Les 6 premiers articles dans l'ordre colex sont 1, 2, 3, 4, 5 et 6. Puisque $t_1 = 0$, il n'y a aucun article dans $\mathcal{B}d^-(\mathcal{F})$.
Au niveau 2, $n_2 = LB_1^{-1}(n_1 - t_1) = LB_1^{-1}(6) = 15$. Puisque $t_2 = 2$, on va mettre les 2-motifs de rang 14 et 15 dans $\mathcal{B}d^-(\mathcal{F})$, soit 46 et 56.
Puis, $n_3 = LB_2^{-1}(n_2 - t_2) = LB_2^{-1}(13) = 13$. Puisque $t_3 = 3$, on insère les 3-motifs de rang 11, 12 et 13 dans $\mathcal{B}d^-(\mathcal{F})$, i.e. 126, 136 et 236.
Enfin, $n_4 = LB_3^{-1}(n_3 - t_3) = LB_3^{-1}(10) = 5$. Parmi ces 5 motifs, on insère le dernier ($t_1 = 1$) dans la bordure négative, soit 2345.
On obtient finalement :
$\mathcal{B}d^-(\mathcal{F}) = \{46, 56, 126, 136, 236, 2345\}$

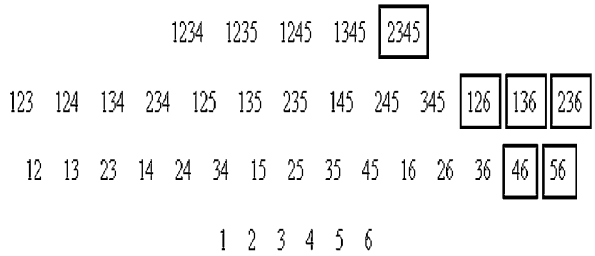

FIG. 5 – Etape 2 : Détermination de la bordure négative (encadrée)

Détermination des motifs maximaux Pour générer la base de transactions voulue de ma-

nière simple, il faut connaître la bordure positive correspondant à la bordure négative générée. Or, les motifs de cette bordure résident forcément avant (toujours dans l'ordre colex) les éléments de la bordure négative, car tous les éléments supérieurs à la bordure négative sont non fréquents.

Le principe est de parcourir les niveaux de haut en bas, du niveau p au niveau 1. Si n_k est le dernier motif de la bordure négative au niveau k, alors les éléments de la bordure positive ont leur rang dans l'intervalle $[1; n_k - t_k]$, qui contient tous les motifs fréquents du niveau. Il reste à déterminer ceux dont tous les sur-ensembles sont non fréquents, c'est à dire ceux qui ne sont pas issus des fréquents du niveau supérieur. Il suffit donc d'exclure les r_k motifs générés par les $k+1$-motifs fréquents, et donc $r_k = LB^1_{k+1}(n_{k+1} - t_{k+1})$. Les motifs de la bordure positive sont ceux dont le rang appartient à l'intervalle $[r_k+1; n_k - t_k]$.

Cette étape est illustrée par l'exemple suivant et la figure 6.

Exemple 10 Reprenons le même exemple. Les 4 premiers 4-motifs dans l'ordre colex, qui sont 1234, 1235, 1245 et 1345, sont dans $\mathcal{B}d^+(\mathcal{F})$, puisque tous leurs sur-ensemble sont considérés non fréquents par construction.
Au niveau 3, les motifs induits par les fréquents du niveau 4 sont au nombre de $r_3 = LB^1_4(4) = 10$. Donc, tous les motifs fréquents du niveau 3 possèdent un sur-ensemble fréquent, et donc ne sont pas dans la bordure positive.
Au niveau 2, on a $r_2 = LB^1_3(10) = 10$ motifs à exclure, il reste donc les trois motifs suivants à insérer dans la bordure positive.
Enfin, $r_1 = LB^1_2(13) = 6$, et donc il ne reste pas d'article à insérer dans la bordure positive. On obtient finalement :
$\mathcal{B}d^+(\mathcal{F}) = \{16, 26, 36, 1234, 1235, 1245, 1345\}$.

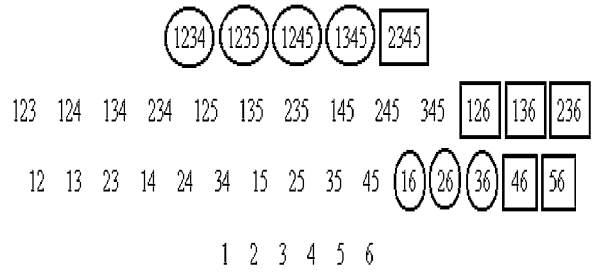

FIG. 6 – Etape 3 : Détermination de la bordure positive (entourée)

La base de transactions résultat peut alors être générée à partir de la bordure positive obtenue.

6 Conclusion

Partant du travail fondateur de [16], cet article explore plus en avant la génération de bases de transactions synthétiques prenant en compte la distribution des bordures positives et négatives des fréquents. En particulier, la nécessité de prendre en compte plus spécifiquement la bordure négative est mise en avant par un théorème et des observations expérimentales. Une méthode, issue de travaux en cours d'approfondissements, est donnée de façon informelle, prenant en entrée une distribution de bordure négative, et fournissant une base données de transactions synthétique "vérifiant" cette bordure. Si la méthode n'est pas encore formellement démontrée, un certain nombre de pistes sont données et permettent de conjecturer de sa correction.

Avant tout, nous pensons que cet article ouvre un grand nombre de perspectives dans ce domaine qui a été peu considéré, malgré l'immense popularité du problème de découverte des motifs fréquents. La méthode proposée pour prendre en compte la bordure négative permet certes d'étudier les performances des algorithmes en fonction de ce paramètre ; toutefois le problème de la prise en compte des distributions relatives des deux

bordures reste toujours ouvert et non trivial. Le problème réside principalement dans le fait de perdre la propriété abondamment utilisée dans cet article, qui est le calcul des motifs induits à l'aide des coefficients binomiaux. En l'état, cette propriété n'est valide que lorsque les motifs des bordures *sont consécutifs dans l'ordre colex* ; il est nécessaire d'assouplir cette contrainte pour réaliser des jeux d'essais plus variés.

Deux autres voix de recherche doivent également être considérées. La première est une généralisation de ce travail, i.e. considérer des bordures de caractéristiques monotones autres que "être fréquent" : les ensembles libres, disjonctifs libres, essentiels. La deuxième est de prendre en compte des caractéristiques importantes des jeux de données, où la notion de bordure ne s'applique pas, comme la distribution des motifs fermés fréquents.

Enfin, nous envisageons l'étude expérimentale elle-même des programmes existants en fonction de ces nouveaux paramètres, conduisant certainement à une nouvelle classification des jeux de données.

Remerciements : Nous remercions Frédéric Flouvat, du laboratoire LIMOS (Clermont-Ferrand), pour avoir mis à notre disposition les distributions des bordures découvertes dans les bases de FIMI.

Références

[1] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *SIGMOD conference, Washington, D.C.*, pages 207–216. ACM Press, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *International Conference on Very Large Data Bases (VLDB'94), Santiago de Chile, Chile*, pages 487–499. Morgan Kaufmann, 1994.

[3] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining frequent patterns with counting inference. *ACM SIGKDD Exploration*, 2(2) :66–75, 2000.

[4] R. J. Bayardo. Efficiently mining long patterns from databases. In L. M. Haas and A. Tiwary, editors, *ACM SIGMOD Conference, Seattle, USA*, pages 85–93, 1998.

[5] B. Bollobás. *Combinatorics.* Cambridge University Press, 1986.

[6] D. Burdick, M. Calimlim, and J. Gehrke. Mafia : A maximal frequent itemset algorithm for transactional databases. In *International Conference on Data Engineering (ICDE'01), Heidelberg, Germany*, pages 443–452. IEEE CS, 2001.

[7] F. Flouvat. Etude expérimentale de la distribution des bordures pour la découverte des motifs fréquents. In *Inforsid 2005, Grenoble, France*, May 2001.

[8] F. Geerts, B. Goethals, and J. Van den Bussche. A tight upper bound on the number of candidate patterns. In *1st IEEE Intl. Conf. on Data Mining*, Nov. 2001.

[9] B. Goethals and M.J. Zaki. Advances in frequent itemset mining implementations : Introduction to fimi03. Technical report, http ://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-90/intro.pdf, 2003.

[10] K. Gouda and M. J. Zaki. Efficiently mining maximal frequent itemsets. In *1st IEEE Intl. Conf. on Data Mining*, San Jose, California, Nov. 2001.

[11] D. Gunopulos, R. Khardon, H. Mannila, S. Saluja, H. Toivonen, and R. S. Sharma. Discovering all most specific sentences. *ACM Transaction on Database System*, 28(2) :140–174, 2003.

[12] J. Han and M. Kamber. *Data Mining : Concepts and Techniques.* Morgan Kaufmann, 2000.

[13] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation : A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1) :53–87, 2004.

[14] G. Katona. A theorem of finite sets. In P. Erdos and G. Katona, editors, *Theory of Graphs*, pages 187–207. Akadémiai Kiadó, Budapest, 1968.

[15] D.-I Lin and Z. M. Kedem. Pincer search : A new algorithm for discovering the maximum frequent set. In H.-J. Schek, F. Saltor, I. Ramos, and G. Alonso, editors, *Extending Database Technology (EDBT'98), Valencia, Spain*, volume 1377 of *Lecture Notes in Computer Science*, pages 105–119. Springer, 1998.

[16] W. A. Maniatty, G. Ramesh, and M. J. Zaki. Feasible itemset distributions in data mining : Theory and application. In *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 284–295. ACM, June 2003.

[17] W. A. Maniatty, G. Ramesh, and M. J. Zaki. Distribution-based synthetic database generation techniques for itemset mining. In *International Database Engineering and Applications Symposium (IDEAS'05), Montreal, Canada*, May 2005.

[18] H. Mannila and H. Toivonen. Multiple uses of frequent sets and condensed representations (extended abstract). In *International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 189–194. AAAI Press, 1996.

[19] H. Mannila and H. Toivonen. Levelwise Search and Borders of Theories in Knowledge Discovery. *Data Mining and Knowledge Discovery*, 1(1) :241–258, 1997.

[20] T. Uno, T. Asai, Y. Uchida, and H. Arimura. An efficient algorithm for enumerating closed patterns in transaction databases. In *Discovery Science*, pages 16–31, 2004.