Vers la génération de jeux de données synthétiques réalistes pour les motifs fréquents

Pascale Bergeret, Frédéric Flouvat, Jean-Marc Petit

Vers la génération de jeux de données synthétiques réalistes pour les motifs fréquents

Pascale Bergeret, Frédéric Flouvat, Jean-Marc Petit
Université de Lyon, INSA-Lyon, LIRIS, UMR5205 CNRS, F-69621, France

Résumé

En fouille de données et plus particulièrement pour les problèmes de motifs fréquents, les algorithmes sont toujours évalués sur des jeux de données afin de montrer leur faisabilité en pratique. Le succès ou l'échec des algorithmes peut parfois s'expliquer par rapport aux caractéristiques des données, par exemple données denses ou éparses. Ces jeux de données peuvent être synthétiques, i.e. générés de façon automatique ou provenir d'applications réelles offrant ainsi un gage d'applicabilité. Paradoxalement, il n'est pas rare que les données synthétiques ne correspondent jamais à des données réelles et inversement, que les données réelles ne puissent pas être représentées par des données synthétiques. Dans ce contexte, c'est la validité même des campagnes de tests qui est posée.

Dans ce papier, nous proposons de générer des jeux de données synthétiques « réalistes » au sens où ils peuvent s'approcher des caractéristiques de n'importe quel jeu de données réel. Pour le problème des motifs fréquents, nous choisissons de définir la « signature » d'un jeu de données par la distribution des bordures positive et négative. À partir d'une structuration fine de l'ordre colex, une proposition théorique est faite permettant d'éloigner la bordure positive de la bordure négative d'une certaine valeur.

Ainsi, un algorithme a pu être conçu puis a été implémenté et testé sur des jeux de données réels confirmant les résultats théoriques. L'intérêt est clairement de pouvoir envisager de véritables campagnes de tests d'algorithmes en s'affranchissant des données réelles.

Mots clés ordre colex, motifs fréquents, bases de données de transactions, bancs d'essais

1 Introduction

La constitution de jeux de données a toujours été une préoccupation majeure en informatique, par exemple en algorithmique, en base de données ou plus récemment en fouille de données. Ces jeux de données peuvent être synthétiques, i.e. générés de façon automatique ou provenir d'applications réelles offrant ainsi un gage d'applicabilité. Néanmoins, le fait de travailler sur des données réelles, quand elles sont disponibles[1], ne garantit pas grand chose pour autant. Paradoxalement, il n'est pas rare que les données synthétiques ne correspondent jamais à des données réelles et inversement, que les données réelles ne puissent pas être représentées par des données synthétiques. Dans ce contexte, c'est la validité même des campagnes de tests qui est posée.

Ces campagnes de tests [FIMI03, FIMI04, OSDM05] permettent d'évaluer plus justement les activités du domaine en mettant à la disposition de tous un environnement de test commun, allant des jeux d'essais diversifiés réels ou synthétiques (provenant du générateur du groupe de recherche d'IBM Almaden) aux codes sources des algorithmes.

On ne peut que constater qu'il est difficile d'échelonner les algorithmes par rapport aux jeux de données et donc de comprendre dans quelles conditions tel type d'algorithme sera performant ou non.

Il est bien clair que les jeux de données ne peuvent pas se caractériser dans l'absolu : il faut leur associer une tâche particulière. Dans ce pa-

[1] Dans le site de FIMI [FIMIRep], il n'existe que 12 jeux de données réels.

pier, nous allons considérer le problème d'énumération de motifs fréquents dans les bases de données de transactions. L'objectif est de proposer une approche de génération de jeux de données synthétiques qui permette de simuler les jeux de données réels tout en offrant un support à la génération de tout jeu de données.

Usuellement, les données sont caractérisées par le nombre d'articles, le nombre de transactions et la taille moyenne des transactions. [RMZ03] a montré qu'un paramètre important semblait être la distribution de la bordure positive des motifs fréquents. Ce constat a motivé la construction de bases de transactions synthétiques dont la bordure positive a une distribution que l'on peut choisir [RMZ03]. Comme montré dans [FMP05], les algorithmes d'énumération de motifs fréquents sont influencés par les distributions des bordures positive et négative caractérisant la solution. Plus précisément, il est montré expérimentalement que la « distance » (symbolisée par d sur la figure 1) entre le pic de la distribution de la bordure positive et celui de la bordure négative est un paramètre prépondérant pour expliquer le comportement des algorithmes. Enfin, il est remarqué dans [DDM05] que pour une même distribution de bordure positive, la distribution de la bordure négative de la base réelle ne ressemble en rien à la distribution de la bordure négative associée à la bordure positive construite par [RMZ03], comme montré sur la Figure 1.

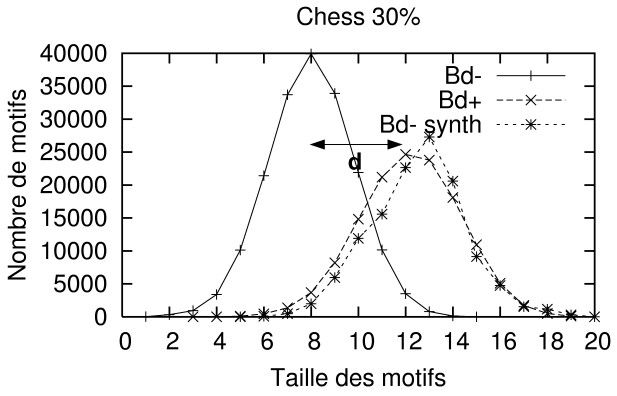

FIG. 1 – Exemple des générations actuelles

De fait, les algorithmes proposés construisent des données pour lesquelles les distributions des bordures positive et négative ont toutes les chances d'être « proches ».

L'hypothèse faite dans cette étude est que la distribution des bordures positive et négative, avec le nombre d'articles doivent permettre de générer tout type de jeu de données.

Contributions L'objectif de ce papier est la génération de jeux de données synthétiques « réalistes » au sens où ils peuvent s'approcher des caractéristiques de n'importe quel jeu de données réel. Pour le problème des motifs fréquents, nous choisissons de définir la « signature » d'un jeu de données par la distribution des bordures positives et négatives. À partir d'une structuration fine de l'ordre colex, une proposition théorique est faite permettant d'éloigner la bordure positive de la bordure négative. Ainsi, un algorithme a pu être conçu partant d'une distribution de bordure négative et d'un facteur d'éloignement avec la distribution de la bordure positive. L'intérêt est clairement de pouvoir envisager de véritables campagnes de tests d'algorithmes en s'affranchissant des données réelles.

L'approche a été implémentée et testée sur des jeux de données réels confirmant les résultats théoriques.

Il faut noter que la génération proprement dite des jeux de données n'est pas abordée dans ce papier. En effet, à partir des distributions des bordures, cette étape est relativement aisée, voir par exemple celle proposée dans [RMZ03] en partant de la bordure positive. Ce point est important au sens où il nous permet de nous détacher du problème particulier des motifs fréquents : de fait, ce que nous proposons dans ce papier peut s'appliquer à tout problème d'énumération de motifs dont la solution peut se représenter par des bordures positive et négative dans un cadre ensembliste (appelé « problèmes représentables par des ensembles dans [MT97] »). Néanmoins, afin de simplifier le discours, nous parlerons de motifs fréquents dans des bases de données de transactions, ce problème étant simple et très étudié.

Organisation du papier Les préliminaires sont donnés en section 2. Une synthèse des tentatives de génération de bordures faites dans [RMZ03] et [DDM05] est proposée dans la section 3. Dans la

section 4 on introduit la notion de bloc de motifs consécutifs dans l'ordre colex en exhibant un exemple mettant en évidence l'importance de ces blocs dans le positionnement de la bordure positive par rapport à la bordure négative. Puis une structure fine de l'ordre colex est mise en relief. La section 5 présente un algorithme de génération de bordures négative et positive pour une distribution de bordure négative donnée et un paramètre entier α. Les résultats des expérimentations menées avec une implémentation de l'algorithme sont donnés dans la section 6. On termine dans la section 7 par les perspectives ouvertes par cette étude.

2 Préliminaires

Définition 1 [Motifs] *Soit (N) l'ensemble des entiers naturels $\{1, 2, \cdots, N\}$. Un élément $X \in (N)$ est appelé un* article. *On note $\mathcal{I} = (N)$ l'ensemble constitué des N articles. Un sous-ensemble non vide de $\mathcal{I}$ est appelé un* motif. *L'ensemble des parties de $\mathcal{I}$, noté $\mathcal{P}(\mathcal{I})$, est l'ensemble de tous les motifs possibles de $\mathcal{I}$. Un motif de cardinal k, $X = \{x_1, \cdots, x_k\}$ est appelé un k-*motif. *On dit aussi que k est la* taille *du motif. Pour simplifier, on omet les accolades et on écrit $X = x_1 \cdots x_k$, les x_i étant par convention écrits du plus petit (x_1) au plus grand (x_k). L'ensemble de tous les k-motifs de $\mathcal{I}$ est noté $\mathcal{I}^{(k)}$.* ▲

Définition 2 [Collections de motifs] *Un ensemble $\mathcal{F} \subseteq \mathcal{P}(\mathcal{I})$ tel que $\emptyset \notin \mathcal{F}$ est appelé une* collection de motifs. *La* k-collection de $\mathcal{F}$, *notée $\mathcal{F}_k$, est la collection de tous les k-motifs de $\mathcal{F}$.*
Soit $f_k = |\mathcal{F}_k|$ le cardinal de $\mathcal{F}_k$. Soit $l \leq N$ la taille du plus long motif de $\mathcal{F}$. La distribution *de $\mathcal{F}$ est la séquence constituée par les cardinaux des ensembles de k-motifs de $\mathcal{F}$: $< \mathcal{F} > = < f_1, \cdots, f_l >$.* ▲

Définition 3 [Bordure positive, bordure négative] *L'ensemble de parties $\mathcal{F}$ est dit* fermé inférieurement *pour $\subseteq$ lorsque pour tout ensemble X dans $\mathcal{F}$, tout sous-ensemble de X est dans $\mathcal{F}$.*
Soit $\mathcal{F}$ un ensemble fermé inférieurement pour $\subseteq$. On note $\mathcal{NF}$ le complémentaire de $\mathcal{F}$ dans $\mathcal{P}(\mathcal{I})$. La bordure positive de $\mathcal{F}$, notée $\mathcal{B}d^+(\mathcal{F})$, est l'ensemble des éléments de $\mathcal{F}$ maximaux pour $\subseteq$. La bordure négative de $\mathcal{F}$, notée $\mathcal{B}d^-(\mathcal{F})$, est l'ensemble des éléments minimaux de $\mathcal{NF}$. L'ensemble $\mathcal{F}$ est entièrement défini par la seule donnée de $\mathcal{B}d^+(\mathcal{F})$ ou de $\mathcal{B}d^-(\mathcal{F})$. De plus, $\mathcal{B}d^+(\mathcal{F})$ est la plus petite collection à partir de laquelle $\mathcal{F}$ peut être déduit. ▲

Définition 4 [Ordres lex et colex] *Soient X et Y deux ensembles de $\mathcal{F}$ tels que X et Y sont distincts. Notons les $X = x_1 \cdots x_k$ and $Y = y_1 \cdots y_k$.*
• *L'ordre lexicographique strict $<_l$ est défini ainsi :*
$X <_l Y$ si et seulement si
$x_1 < y_1$ OU il existe $1 < \kappa \leq k$ tel que
$([\forall i, 1 \leq i < \kappa \Rightarrow x_i = y_i]$ ET $x_\kappa < y_\kappa)$.
• *L'ordre colex strict $<_C$ est défini ainsi :*
$X <_C Y$ si et seulement si
$x_k < y_k$ OU il existe $1 \leq \kappa < k$ tel que
$([\forall i, \kappa < i \leq k \Rightarrow x_i = y_i]$ ET $x_\kappa < y_\kappa)$.
La relation lexicographique au sens large $\leq_l$ et la relation colex au sens large $\leq_C$ sont des relations d'ordre total sur $\mathcal{I}^{(k)}$.
On définit le rang *d'un k-motif comme étant sa position dans la liste des k-motifs ordonnés, le premier k-motif ayant le rang* 1. ▲

Exemple 1 *1236 $<_l$ 2345 mais 2345 $<_C$ 1236 ; 1256 $<_l$ 1346 mais 1346 $<_C$ 1256* ♢

Énumération dans l'ordre colex Le premier motif est $12 \cdots k$. Pour générer le successeur x' d'un motif $x = x_1 \cdots x_k$:
(a) On note t le plus petit des indices i de $\{1, \cdots, k-1\}$ pour lesquels $x_{i+1} > x_i + 1$. Si tous les entiers du motif sont consécutifs, on prend t égal à k.
(b) On pose $x'_t = x_t + 1$.
(c) Pour $i < t$, on pose $x'_i = i$. Pour $i > t$, on pose $x'_i = x_i$.

Exemple 2 *Liste des motifs composés de 3 articles en ordre colex :* 123, 124, 134, 234, 125, 135, 235, 145, 145, 245, 345, ... ♢

Intérêt de l'ordre colex Le premier motif de taille k faisant intervenir l'entier l n'apparait qu'après l'énumération de tous les motifs faisant intervenir uniquement les articles 1, 2, ..., $l-1$. Une conséquence intéressante est qu'un k-motif possède toujours le même rang quelque soit le nombre d'articles, ce qui n'est pas le cas pour l'ordre lexicographique.

Exemple 3 *Le motif* 134 *apparait toujours en position* 3 *dans l'énumération en ordre colex. Dans l'ordre lexicographique,* 134 *apparait en position* 3 *lorsqu'il y a* 4 *articles mais en position* 4 *lorsqu'il y a 5 articles.* ♢

Définition 5 [Motif fréquent dans une base de transactions] [AIS93] *Une base de transactions est la donnée d'un triplet* $(\mathcal{T}, \mathcal{I}, \mathcal{R})$ *où* $\mathcal{T}$ *est un ensemble fini dit ensemble de transactions,* $\mathcal{I}$ *un ensemble d'articles et* $\mathcal{R}$ *une partie de* $\mathcal{T} \times \mathcal{I}$*. Étant donné un réel positif dans l'intervalle* $[0, 1]$ *appelé seuil, un motif est dit* fréquent pour ce seuil *lorsque la proportion des transactions qui le contiennent est supérieure à ce seuil.* ▲

Dans cet article, la notion de fréquent n'est abordée qu'au travers de la propriété d'anti-monotonie du prédicat fréquent ou encore de monotonie du prédicat non-fréquent :

Propriété 1 *[AIS93] L'ensemble des motifs fréquents pour un seuil donné est un ensemble fermé inférieurement, c'est-à-dire que* **tout sous-motif d'un motif fréquent est fréquent** *ou encore* **tout sur-motif d'un motif non fréquent est non fréquent**. ■

Par conséquent, on peut parler de bordure négative et bordure positive des fréquents.

3 Travaux relatifs et critiques

Il existe à notre connaissance trois tentatives de génération de bases de transactions synthétiques. La première est celle du groupe de recherche d'IBM Almaden Quest utilisée pour des bases dont on peut trouver quelques exemplaires sur le site du FIMI [FIMIRep]. Le peu de ressemblance avec des données réelles a motivé l'étude faite dans [RMZ03] dans laquelle un nouveau paramètre essentiel est pris en compte : la distribution de la bordure positive des fréquents. Il est montré que pour toute distribution, il existe une bordure positive de fréquents ayant cette distribution. Dans [DDM05], le paramètre pris en compte est la distribution de la bordure négative des fréquents. Il est à noter qu'une distribution donnée n'est pas la distribution d'une seule bordure possible. Les méthodes de construction d'une bordure de distribution donnée exposées dans [RMZ03] et [DDM05] utilisent le plus petit nombre d'articles possible, en se basant sur l'ordre colex pour la génération. La propriété suivante de l'ordre colex est exploitée :

Propriété 2 de l'ordre colex [RMZ03]

Étant donnés deux entiers naturels l *et* h*, la collection de* l*-motifs et de cardinal* h *qui induit le plus petit nombre de* $(l-1)$*-motifs est la collection des* h *premiers motifs dans l'ordre colex.* ***Les*** j***-motifs induits dans les niveaux inférieurs sont les premiers*** j***-motifs consécutifs dans l'ordre colex****. Tout autre collection de* l*-motifs et de cardinal* h *a un nombre de* j *motifs induits plus grand ou égal.* ■

On rappelle ci-dessous le procédé de génération de la bordure positive des fréquents proposé dans [RMZ03] puis celui de la bordure négative proposé dans [DDM05]. On explicite ensuite pourquoi l'un et l'autre ne peuvent en général pas rendre compte de cas mentionnés dans l'introduction où les distributions des bordures positive et négative ne sont pas « proches » l'une de l'autre (cf. figure 1, lorsque la distance entre les pics des distributions est importante).

Génération d'une bordure positive de distribution donnée selon [RMZ03] Étant donnée $S^+ =< s_1, \cdots, s_l >$, on construit par récurrence la collection $\mathcal{B}d^+$ de fréquents suivante :

- Initialisation : on met dans la collection les s_l premiers l-motifs dans l'ordre colex.
- Pour k dans $\{l-1, \cdots, 1\}$, on ajoute à la collection les s_k k-motifs consécutifs dans l'ordre colex qui ont pour rang $r_k + 1, \ldots r_k + s_k$, où r_k est le rang du dernier k-motif induit par les motifs fréquents de taille supérieure.

Il est par ailleurs possible de déterminer indépendamment de la construction effective de la bordure le nombre d'articles nécessaires par un calcul utilisant des cœfficients binomiaux [RMZ03].

Génération d'une bordure négative de distribution donnée selon [DDM05] Étant donné $S^- =< t_1, \cdots, t_p >$, on détermine tout d'abord par récurrence le nombre r'_1 d'articles qui seraient nécessaires à la construction d'une bordure positive de distribution S^- selon [RMZ03]. Puis, en utilisant un nombre d'articles égal à r'_1, on procède à

une construction par récurrence dans l'ordre colex inverse, du niveau le plus petit au niveau le plus grand, de la collection $\mathcal{B}d^-$ suivante :

• Au premier niveau, on insère dans $\mathcal{B}d^-$ les t_1 derniers 1-motifs consécutifs dans l'ordre colex, c'est-à-dire ceux de rang $r'_1 - t_1 + 1, \ldots, r'_1$. Ces t_1 1-motifs non fréquents induisent des 2-motifs non fréquents qui sont les derniers 2-motifs consécutifs dans l'ordre colex à partir d'un certain rang $r'_2 + 1$. Autrement dit, les r'_2 premiers 2-motifs dans l'ordre colex et de rang compris entre 1 et r'_2 ont tous leurs 1-motifs induits fréquents.

• À chaque niveau $k \geq 2$, à partir d'un certain rang $r'_k + 1$, les derniers k-motifs consécutifs dans l'ordre colex sont non fréquents, en tant que surmotifs des $(k-1)$-motifs non fréquents de rang supérieur ou égal à $r'_{k-1} - t_{k-1} + 1$. Autrement dit, r'_k est le cardinal de l'ensemble des premiers k-motifs dans l'ordre colex dont tous les $(k-1)$-motifs induits sont fréquents. On insère dans $\mathcal{B}d^-$ les t_k k-motifs consécutifs dans l'ordre colex dont le rang est compris entre $r'_k - t_k + 1$ et r'_k.

Défaut des deux constructions Les constructions de [RMZ03] et [DDM05] utilisent un nombre minimum d'articles. L'inconvénient majeur est que le choix de prendre à chaque niveau des motifs consécutifs dans la bordure fige sensiblement la position relative des distributions. En effet, si pour simplifier le discours, les distributions des deux bordures sont non nulles aux niveaux considérés, les constructions proposées dans [RMZ03] et dans [DDM05] donnent pour chaque niveau (cf. fig. 2)

- r_k motifs consécutifs dans l'ordre colex et qui sont fréquents, suivis de
- s_k motifs consécutifs dans l'ordre colex qui composent le niveau k de la bordure positive, suivis de
- t_k motifs consécutifs dans l'ordre colex qui forment le niveau k de la bordure négative,
- suivis de motifs consécutifs dans l'ordre colex et qui sont non fréquents.

Le dernier motif fréquent $x_1 \cdots x_k$ de la bordure positive $\mathcal{B}d^+{}_k$ au niveau k induit en particulier le dernier motif $x_2 \cdots x_k$ de la plage des fréquents $\mathbf{F}_{k-1}$ du niveau $k-1$. Le premier motif non fréquent de la plage $\mathbf{NF}_k$ des non fréquents est induit par le premier motif $x'_1 \cdots x'_{k-1}$ de la bordure négative $\mathcal{B}d^-{}_{k-1}$ du niveau $k-1$. Les cardinaux de $\mathcal{B}d^+{}_{k-1}$ et de $\mathcal{B}d^-{}_k$ évoluent de la même façon. Lorsque l'un augmente, l'autre augmente ou reste à la même valeur (fig. 2).

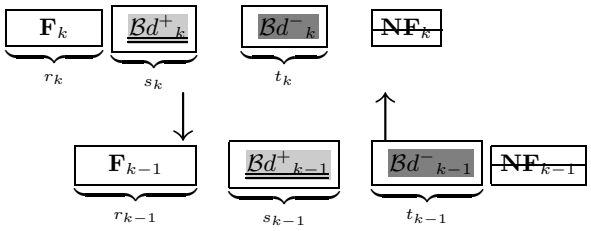

FIG. 2 – Forte dépendance entre $|\mathcal{B}d^-{}_k|$ et $|\mathcal{B}d^+{}_{k-1}|$ dans [RMZ03] et [DDM05]

Cette forte dépendance explique les observations faites dans [DDM05] concernant les positions relatives des bordures positive et négative synthétiques. La construction proposée de bordure négative de fréquents n'est pas un remède au non réalisme des bases de transactions synthétiques de [RMZ03]. La construction de [DDM05] souffre du même défaut. Si l'on tente de construire une base de transactions à partir de distributions réalistes ou réelles, on a le choix entre « coller » à la bordure positive ou « coller » à la bordure négative en étant éloigné dans les deux cas de la bordure duale réelle (cf. fig. 1).

4 Blocs de motifs consécutifs et propriétés

Nous nous intéressons aux problèmes soulevés dans la section précédente et nous introduisons la notion de blocs de motifs consécutifs sur un exemple.

4.1 Exemple introductif

On examine une construction selon la méthode de [DDM05] sur l'exemple très particulier où $\mathcal{S}^- =< 0, 0, 20 >$ (distribution non nulle sur le seul niveau 3). On s'intéresse à la manière dont les motifs sont intégrés à la bordure positive à mesure de l'intégration dans l'ordre colex inverse des motifs dans la bordure négative. La figure 3 de la page 6 présente l'état des bordures à chaque nouvelle intégration d'éléments dans la bordure positive. Pour la distribution $\mathcal{S}^-$ donnée, le nombre nécessaire d'ar-

ticles est six. La construction de la bordure négative commence par l'intégration du dernier motif 456 du niveau 3. Ceci ne provoque aucune entrée de motif dans la bordure positive. Le premier motif à entrer dans la bordure positive est 56 ; il n'y entre que lorsque les quatre sur-motifs 156, 256, 356, 456 sont intégrés à la bordure négative. Pour que le motif 46 entre dans la bordure positive, il faut ensuite intégrer les sur-motifs 146, 246 et 346 dans la bordure négative car le sur-motif 456 est déjà non fréquent. Pour que le motif 36 entre à son tour dans la bordure positive, l'intégration des deux sur-motifs 136 et 236 dans la bordure négative est nécessaire. Puis l'intégration du seul motif 126 dans la bordure négative fait entrer en même temps les deux motifs 16 et 26 dans la bordure positive. La bordure négative est finalement constituée de tous les motifs de trois articles. La bordure positive est constituée de tous les motifs de deux articles. On voit que le cardinal de la bordure positive augmente chaque fois qu'est intégré à la bordure négative un certain ensemble de 3-motifs consécutifs dans l'ordre colex qui ont leur 2-motif suffixe identique. Ces ensembles de motifs consécutifs ayant même suffixe apparaissent sur la figure 3 regroupés dans des boîtes. On introduit le terme de « *blocs* » pour les désigner. La notion de bloc est développée en détail dans la section 4.2.

Ces blocs jouent un rôle important. C'est l'intégration d'un bloc complet dans la bordure négative a un niveau k (dans l'exemple k vaut 3) qui amène l'intégration d'un ou plusieurs motifs du niveau $k-1$ dans la bordure positive. Imposer que les motifs de la bordure négative soient consécutifs dans l'ordre colex amène l'intégration de blocs complets et implacablement une bordure positive correspondante au niveau immédiatement inférieur. **Pour « séparer » les bordures, il est donc nécessaire de ne pas prendre les blocs entiers dans la bordure négative**.

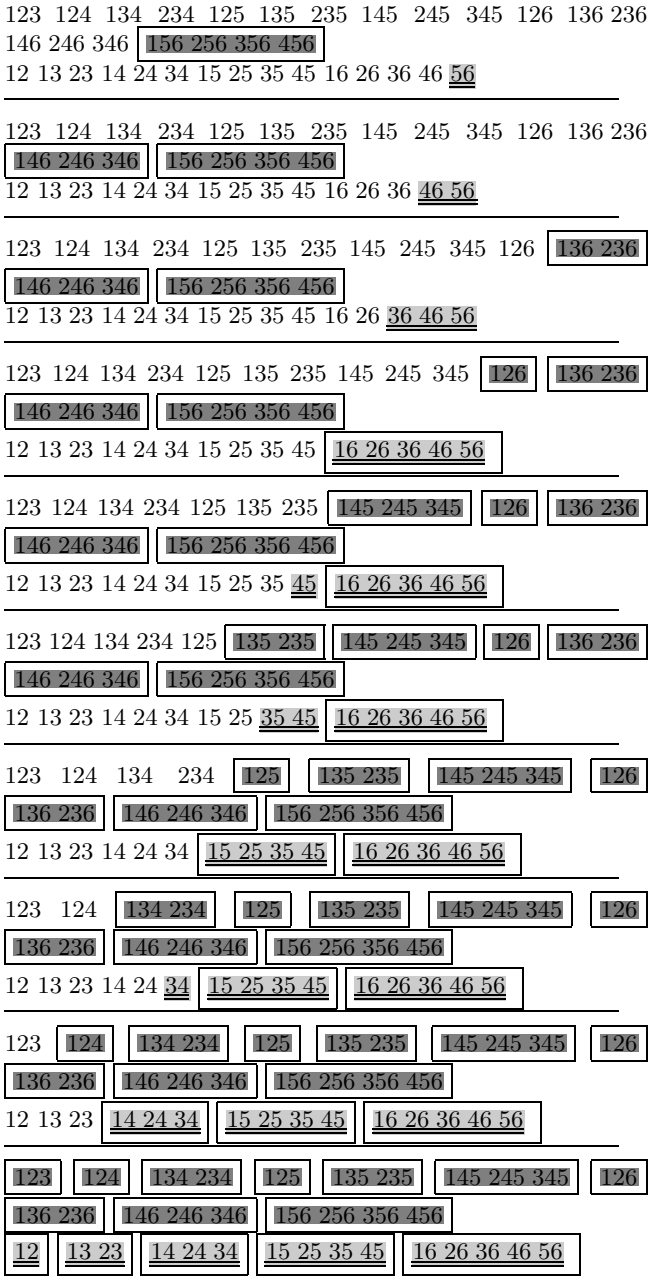

FIG. 3 – Légende : bordure négative, bordure positive. Évolution des bordures lors de la construction d'une bordure négative de distribution $< \mathcal{B}d^- > = < 0, 0, 20 >$ selon [DDM05]. Seuls les niveaux utiles 2 et 3 sont présentés. De haut en bas apparaissent chacune des étapes amenant de nouveaux motifs dans $\mathcal{B}d^+$. Cet exemple met en évidence que l'intégration d'un motif ou quelquefois plusieurs motifs dans la bordure positive correspond à l'intégration dans la bordure négative d'un « bloc » de sur-motifs consécutifs dans l'ordre colex. La solution apparait dans le « pavé » du bas.

4.2 Notion de blocs dans l'énumération colex

L'exemple exposé dans 4.1 a montré l'utilité d'introduire la notion de blocs de motifs consécutifs. On développe ici cette notion. Puis on donne ensuite des propriétés sur les blocs.

Caractérisation d'un bloc.

On peut décomposer un niveau k en une suite contigüe de blocs contenant des k-motifs successifs dans l'ordre colex ayant même suffixe de taille $(k-1)$. Un bloc du niveau k est caractérisé par le $(k-1)$-motif qui constitue le suffixe de chaque k-motif du bloc. Aucun autre bloc ne comporte ce suffixe. On choisit en fait de caractériser le bloc par son dernier élément. Par exemple $\mathcal{B}_{4579}$ est le bloc $\boxed{1579\ 2579\ 3579\ 4579}$ ou plus généralement :

Définition 6 *Soient k un indice et x_2 un entier tels que $k \geq 2$ et $x_2 \geq 2$.*
Pour $x_2 < \cdots < x_k$, l'ensemble des k-motifs ayant même $(k-1)$-suffixe $x_2 \cdots x_k$ est composé des k-motifs consécutifs dans l'ordre colex compris entre $1x_2 \cdots x_k$ et $(x_2-1)x_2x_3 \cdots x_k$. On nomme cet ensemble **bloc $\mathcal{B}_{(x_2-1)x_2 \cdots x_k}$**.
Dans l'énumération colex, on le fera apparaitre entouré d'un rectangle :
$\boxed{1x_2x_3 \cdots x_k \ \ 2x_2x_3 \cdots x_k \ \ \ldots\ldots \ \ (x_2-1)x_2x_3 \cdots x_k}$.
Son cardinal est (x_2-1). On dit aussi que la **taille du bloc $\mathcal{B}_{(x_2-1)x_2 \cdots x_k}$** *est (x_2-1).* ▲

Définition 7 **Rang d'un motif à l'intérieur de son bloc** *Soit un k-motif $x_1x_2 \cdots x_k$. Il appartient au bloc de taille $\mathcal{B}_{(x_2-1)x_2 \cdots x_k}$ de taille x_2-1. Son rang dans le bloc est x_1.* ▲

4.3 Blocs de premiers sur-motifs

L'exemple donné dans la section 4.1 met en évidence que l'intégration d'un motif ou quelquefois plusieurs motifs dans la bordure positive correspond à l'intégration dans la bordure négative d'un « bloc » de sur-motifs consécutifs dans l'ordre colex. C'est pourquoi on introduit la notion de blocs de premiers sur-motifs. Elle va permettre de formaliser les relations entre deux niveaux consécutifs de motifs. Afin de simplifier les écritures, on définit auparavant l'indice de croissance colex. On s'intéresse ensuite aux premiers sur-motifs des motifs d'un bloc de taille supérieure ou égale à 3, puis aux premiers sur-motifs des blocs de taille 1 et 2.

Définition 8 **Indice de croissance colex d'un motif** *Soit $x = x_1 \cdots x_k$ un k-motif de longueur $k \geq 2$. On note J le plus petit indice compris entre 1 et $k-1$ pour lequel on a $x_J + 1 < x_{J+1}$. Si un tel indice n'existe pas, c'est-à-dire si tous les entiers du motif sont consécutifs, on prend $J = k$. On dit que J est l'***indice de croissance colex de x***.* ▲

Cette dénomination vient du fait que c'est l'entier x_J qui est incrémenté de 1 lorsqu'on calcule le successeur dans l'ordre colex du motif $x_1 \cdots x_k$ (cf. étape (b) de l'énumération colex page 3). Par exemple, l'indice de croisance colex du motif 2346 est 3 et son successeur est 1256.

Définition 9 **Bloc de premiers sur-motifs d'un motif $x_1 \cdots x_k$ d'un bloc de taille supérieure ou égale à 3.**
• **Cas $x_1 \geq 2$.** *Soit $x_1 \cdots x_k$ un k-motif ayant un rang x_1 supérieur ou égal à 2 dans un bloc de taille supérieure ou égale à 3. Ses (x_1-1) premiers sur-motifs ont tous le même suffixe $x_1x_2 \cdots x_k$ et le sur-motif suivant a un suffixe qui ne commence pas par x_1. Les (x_1-1) premiers sur-motifs appartiennent donc au bloc $\mathcal{B}_{(x_1-1)x_1x_2 \cdots x_k}$ de taille (x_1-1) du niveau $k+1$.* **Le bloc $\mathcal{B}_{(x_1-1)x_1x_2 \cdots x_k}$ est appelé bloc de premiers sur-motifs du motif $x_1 \cdots x_k$.**
• **Cas $x_1 = 1$.** *Le premier sur-motif de $1x_2 \cdots x_k$ est $12x_2 \cdots x_k$. Il appartient à un bloc de taille 1, qui est déjà le bloc de premiers sur-motifs du motif $2x_2 \cdots x_k$.* **Le bloc $\mathcal{B}_{12x_2 \cdots x_k}$ est encore appelé bloc de premiers sur-motifs du motif $1x_2 \cdots x_k$.** ▲

Exemple 4 *Les motifs* 168, 268, 368, 468, 568 *du bloc $\mathcal{B}_{568}$* $\boxed{168\ 268\ 368\ 468\ 568}$ *ont pour bloc de premiers sur-motifs les blocs indiqués ci-dessous.*
168 *:* $\mathcal{B}_{1268}$ $\boxed{1268}$*;* 268 *:* $\mathcal{B}_{1268}$ *encore ;* 368 *:* $\mathcal{B}_{2368}$ $\boxed{1368\ 2368}$*;* 468 *:* $\mathcal{B}_{3468}$ $\boxed{1468\ 2468\ 3468}$*;* 568 *:* $\mathcal{B}_{4568}$ $\boxed{1568\ 2568\ 3568\ 4568}$.

♢

Blocs de premiers sur-motifs des motifs des blocs de taille 1 et 2 On s'intéresse à présent aux sur-motifs des blocs de taille 2 du niveau k, où $k \geq 3$. Pour $k = 2$, il existe un unique bloc de taille 2 qui est $\boxed{13\ 23}$. Pour $k \geq 3$, un bloc de taille 2 du niveau k est de la forme $\boxed{13x_3 \cdots x_k\ 23x_3 \cdots x_k}$ où $x_3 \geq 4$. Il est forcément précédé du bloc de taille 1 : $\boxed{12x_3 \cdots x_k}$. Le premier sur-motif de chacun des trois k-motifs est identique : $\boxed{123x_3 \cdots x_k}$. En fait, le bloc $\boxed{13x_3 \cdots x_k\ 23x_3 \cdots x_k}$ peut être précédé de

plus d'un bloc de taille 1.

Proposition 1 *Le nombre de blocs de taille* 1 *précédant le bloc* $\boxed{13x_3\cdots x_k\ 23x_3\cdots x_k}$ *est* $(J-1)$*, où* J *est l'indice de croissance colex du motif* $23x_3\cdots x_k$*. Ce nombre est toujours plus grand ou égal à 1.*
Pour $J=2$*, un seul bloc de taille* 1 *précède* $\mathcal{B}_{23x_3\cdots x_k}$ *:* $\boxed{12x_3\cdots x_k}$.
Pour $J\geq 3$*, les blocs de taille* 1 *précédant* $\mathcal{B}_{23x_3\cdots x_k}$ *sont les* $(J-1)$ *blocs consécutifs* $\boxed{123\cdots Jx_{J+1}\cdots x_k}$ *à* $\boxed{12x_3\cdots x_k}$. ■

Remarque 1 Les indices de croissance colex des motifs $12\cdots Jx_{J+1}\cdots x_k$ à $23x_3\cdots x_k$ varient ainsi : J, $J-1$, ...,1, J.

Exemple 5 *Pour* $\mathcal{B}_{234578}$*, on écrit des motifs prédécesseurs en dessinant les blocs de taille 1 :*
134568 234568 $\boxed{123478}\boxed{123578}\boxed{124578}\boxed{134578\ 234578}$.
L'indice de croissance colex du motif caractérisant $\mathcal{B}_{234578}$ *est* $J=4$*. Ce bloc est bien précédé de trois blocs de taille* 1*. Le premier bloc de taille 1 est* $\mathcal{B}_{123478}$*. Le prédécesseur de* 123478 *est* 234568 *qui ne constitue pas un bloc de taille 1. L'indice de croissance colex des motifs mis en jeu varie de la façon suivante : 4, 3, 2, 1, 4.* ♢

Définition 10 **Bloc de premier sur-motif des motifs des blocs de taille** 2 **et** 1**.**
Soit J *l'indice de croissance colex du motif* $23x_3\cdots x_k$*. Les deux* k*-motifs du bloc* $\mathcal{B}_{23x_3\cdots x_k}$ *et les* $J-1$ *motifs des blocs de taille* 1 *précédents ont le même premier sur-motif* $123x_3\cdots x_k$*. Le bloc* $\mathcal{B}_{123x_3\cdots x_k}$ *est appelé bloc de premier sur-motif des motifs* $123\cdots Jx_{J+1}\cdots x_k$ *à* $2x_3\cdots x_k$*.* ▲

Exemple 6 *Les 5 motifs des blocs de l'exemple 5 ont le même bloc* $\mathcal{B}_{1234578}$ *de premiers sur-motifs et il est de taille* 1*.* ♢

Enchaînement des blocs

Proposition 2 *Les blocs de premiers sur-motifs des* k*-motifs d'un bloc de taille supérieure ou égale à* 3 *sont consécutifs, c'est-à-dire que les* (x_2-2) *blocs* $\mathcal{B}_{12x_2\cdots x_k},\ldots,\mathcal{B}_{(x_2-2)(x_2-1)x_2\cdots x_k}$ *sont consécutifs, dans cet ordre.* ■

Exemple 7 *Les motifs* 168*,* 268*,* 368*,* 468*,* 568 *du bloc* $\mathcal{B}_{568}$ $\boxed{168\ 268\ 368\ 468\ 568}$ *de l'exemple 4 ont leur bloc de premiers sur-motifs qui s'enchaînent dans l'ordre colex :*

$\boxed{1268}\boxed{1368\ 2368}\boxed{1468\ 2468\ 3468}\boxed{1568\ 2568\ 3568\ 4568}$ ♢

Plus généralement :

Proposition 3 *Soient* X *et* X' *deux motifs de rang respectif* r *et* r' *dans leur bloc et tels que* X' *est le successeur de* X *dans l'ordre colex. Le bloc de premiers sur-motifs de* X' *est le successeur du bloc de premiers sur-motifs de* X*, sauf dans les cas où* (r,r') *vaut* $(1,1)$ *ou* $(1,2)$ *pour lequel les deux blocs de premiers sur-motifs sont confondus.* ■

4.4 Structuration de l'ordre colex

Afin de faciliter la manipulation des motifs des blocs de premiers sur-motifs lorsqu'on s'intéresse aux sur-motifs sur plusieurs niveaux d'affilée, on définit l'application Θ qui à tout motif de taille k fait correspondre l'ensemble de ces premiers sur-motifs de taille $k+1$ (déf. 9 et 10). On suit ainsi facilement la propagation des non fréquents sur les niveaux supérieurs.

Par ailleurs, lorsqu'on construit un exemple de bordure négative avec sa bordure positive correspondante, on énumère habituellement tous les motifs de chaque niveau (cf figure 3). Afin de mieux expliquer, on utilisera une nouvelle représentation dite « *représentation minimaliste* ».
Représentation minimaliste On ne fait apparaître qu'un rectangle symbolisant chaque motif. Ainsi, on met en évidence les blocs qui s'enchaînent en regroupant les rectangles et l'application Θ en traçant un trait entre un motif et son bloc de premiers sur-motifs au niveau juste au dessus. Le lecteur peut d'ores et déjà se reporter à la figure 4 page 11 utilise cette représentation très pratique (sur de petits exemples) pour comprendre ce qui fait marcher la construction.

On donne ci-dessous des définitions précises pour Θ et ses itérées.

Définition 11 **Application Θ « Ensemble de premiers sur-motifs de ... » Image d'un motif**
Soit $X=x_1\cdots x_k$ *un motif de taille* $k\geq 2$ *et* J *son indice de croissance colex. Le motif* X *appartient à un bloc de taille* $p=x_2-1$*. On note* $\mathcal{B}_{p(p+1)x_3\cdots x_k}$ *le bloc auquel* X *appartient. On pose :*

si $\boldsymbol{p \geq 2}$ *et* $x_1 = 1$, $\Theta(X) = \mathcal{B}_{12(p+1)x_3\cdots x_k}$;
si $\boldsymbol{p \geq 2}$ *et* $x_1 \geq 2$, $\Theta(X) = \mathcal{B}_{(x_1-1)x_1(p+1)x_3\cdots x_k}$;
si $\boldsymbol{p = 1}$, $\Theta(X) = \mathcal{B}_{12\cdots J(J+1)x_{J+1}\cdots x_k}$.

Image d'un ensemble $\mathcal{E}$ de motifs

On note encore $\Theta(\mathcal{E})$ *l'ensemble des premiers sur-motifs de chacun des motifs* e_i *de* $\mathcal{E}$ *:*
$\Theta(\mathcal{E}) = \bigcup_{e_i \in \mathcal{E}} \Theta(e_i)$. ▲

Remarque 2 Si on note b^1 et b^2 les deux premiers motifs d'un bloc de taille $p \geq 2$, on a $\Theta(b^1) = \Theta(b^2)$.

Définition 12 **Itérées de Θ** *Soit un ensemble* $\mathcal{E}$ *de motifs. On pose* $\Theta^0(\mathcal{E}) = \mathcal{E}$ *et pour* γ *entier tel que* $\gamma \geq 1$, $\Theta^{\gamma+1}(\mathcal{E}) = \Theta\left(\Theta^{\gamma}(\mathcal{E})\right)$. ▲

Définition 13 $\boldsymbol{\Theta^{-1}}$
• *Soit* $\mathcal{B}_{p(p+1)x_3\cdots x_k}$ *un bloc de taille* $p \geq 2$. *On note* $\Theta^{-1}(\mathcal{B}_{p(p+1)x_3\cdots x_k})$ *le singleton composé du motif de taille* $k-1$ *qui a* $\mathcal{B}_{p(p+1)x_3\cdots x_k}$ *comme bloc de premiers sur-motifs :*
$\Theta^{-1}(\mathcal{B}_{p(p+1)x_3\cdots x_k}) = \{(p+1)x_3\cdots x_k\}$
• *Soit* $\mathcal{B}_{12x_3\cdots x_k}$ *un bloc de taille* 1 *et soit* J *l'indice de croissance colex de* $12x_3\cdots x_k$.
On note $\Theta^{-1}(\mathcal{B}_{12x_3\cdots x_k})$ *l'ensemble des* J *motifs de taille* $k-1$ *et consécutifs dans l'ordre colex ayant* $\mathcal{B}_{12x_3\cdots x_k}$ *comme bloc de premier sur-motif, i.e. l'ensemble* $\{12\cdots(J-1)x_{J+1}\cdots x_k, \ldots, 2x_3\cdots x_k\}$ ▲

4.5 Sur-motifs autres que les premiers sur-motifs

Proposition 4 *Soit* $x_1\cdots x_k$ *un motif tel que* $x_1 \geq 2$.
Tous les sur-motifs de $x_1\cdots x_k$ *qui ne sont pas dans le bloc* $\mathcal{B}_{(x_1-1)x_1x_2\cdots x_k}$ *des* (x_1-1) *premiers sur-motifs sont des éléments de rang* x_1 *dans un bloc comportant des motifs plus grands dans l'ordre colex que* $(x_1-1)x_1x_2\cdots x_k$. ■

On peut écrire une proposition du même type dans le cas où $x_1 = 1$ [Ber07]. C'est sur ces propositions que reposent en partie la validité de l'algorithme.

4.6 Autres propriétés concernant les blocs

L'algorithme proposé dans la section suivante manipule des blocs. Les propriétés données ci-dessous seront utilisées pour passer du bloc que l'on traite au bloc suivant dans le traitement.

Proposition 5 **Premier bloc de taille supérieure ou égale à p précédant un bloc de taille p.**
Soit p *tel que* $2 \leq p$ *et soit* $\mathcal{B}_{p(p+1)x_3\cdots x_k}$ *un bloc de taille* p. *On note* J *l'indice de croissance colex du motif* $p(p+1)x_3\cdots x_k$.
Cas $\boldsymbol{J < k}$.
Le premier bloc de taille supérieure ou égale à p *précédant le bloc* $\mathcal{B}_{p(p+1)x_3\cdots x_k}$ *est le bloc* $\mathcal{B}_Z$ *où* Z *est le prédécesseur dans l'ordre colex de* $12\cdots Jx_{J+1}\cdots x_k$.
Le sur-motif du niveau $k+p-1$ *le plus petit dans l'ordre colex de chacun des motifs compris entre* $12\cdots Jx_{J+1}\cdots x_k$ *et* $p(p+1)x_3\cdots x_k$ *est identique, égal à* $12\cdots(p-1)p(p+1)x_3\cdots x_k$.
Cas $\boldsymbol{J = k}$.
Le bloc $\mathcal{B}_{p(p+1)\cdots(p+k-1)}$ *est le premier bloc de taille* p *dans l'ordre colex. Il n'y a donc pas de bloc prédécesseur de taille supérieure ou égale.*
Le sur-motif du niveau $k+p-1$ *le plus petit dans l'ordre colex de chacun des motifs compris entre* $12\cdots k$ *et* $p(p+1)\cdots(p+k-1)$ *est identique, égal à* $1\cdots(p+k-1)$. ■

Proposition 6 **Premier bloc de taille supérieure à $p+1$ prédécesseur d'un bloc de taille p**
Soit p *tel que* $1 \leq p \leq (N-k)$ *et soit* $\mathcal{B}_{p(p+1)x_3\cdots x_k}$ *un bloc de taille* p.
On note J *l'indice de croissance colex du motif* $p(p+1)x_3\cdots x_k$ *et on suppose que* $J < k$.
On pose alors
$W = (p+1)(p+2)\cdots(p+J)x_{J+1}\cdots x_k$ *et*
H *l'indice de croissance colex du motif* W.
Le bloc $\mathcal{B}_W$ *est le premier bloc de taille* $p+1$ *successeur de* $\mathcal{B}_X$ *dans l'ordre colex d'où :*
Cas $\boldsymbol{H < k}$.
Le premier bloc de taille supérieure ou égale à $p+1$ *précédant le bloc* $\mathcal{B}_{p(p+1)x_3\cdots x_k}$ *est le bloc* $\mathcal{B}_Z$ *où* Z *est le prédécesseur de* $1\cdots Hx_{H+1}\cdots x_k$.
Cas $\boldsymbol{H = k}$.

Le bloc $\mathcal{B}_W$ est le premier bloc de taille $p+1$ dans l'ordre colex. Il n'y a donc pas de bloc prédécesseur de taille supérieure ou égale à $p+1$ pour le bloc $\mathcal{B}_{p(p+1)x_3\cdots x_k}$. ■

5 Algorithme

Dans cette section, on décrit page 10 et 12 un algorithme conçu dans l'optique de construire une bordure négative générée au plus prêt d'une distribution de bordure donnée en entrée et dont la bordure positive est « éloignée », l'éloignement étant contrôlé par un paramètre α constant.
La validité de l'algorithme repose sur les propositions de la section 4 et d'autres propositions non données ici mais sur lesquelles sont calqués les traitements proposés. Les démonstrations se trouvent dans [Ber07]. Néanmoins, grâce à la représentation minimaliste mais très visuelle mentionnée dans la section 4.4, les exemples donnés ici permettent de comprendre pourquoi la construction fonctionne sans se plonger dans les calculs.

5.1 Vue d'ensemble de l'algorithme

Les ***entrées de l'algorithme*** sont une **distribution $\mathcal{S}^-$**, un entier **α** représentant l'éloignement.

Les ***sorties de l'algorithme*** sont la **bordure négative $\mathcal{BD}^-$** et **la bordure positive $\mathcal{BD}^+$**, **leur distribution** $\mathcal{D}^-$ et $\mathcal{D}^+$ respectivement, ainsi que **le nombre d'articles** N.

L'algorithme construit une bordure négative $\mathcal{BD}^-$ dont la distribution $\mathcal{D}^-$ est calculée et approche la distribution donnée $\mathcal{S}^-$ par valeurs supérieures, au plus près que le permet la stratégie de construction utilisée (le sens exact de « approche » est donné par le test d'arrêt de la construction de $\mathcal{D}^-$ sur un niveau et exposé à la section 5.2). Le nombre d'articles N qui permet d'approcher au mieux la distribution donnée $\mathcal{S}^-$ est calculé. En effet, l'algorithme principal est une boucle sur N qui incrémente N de 1 à partir de $N_0 = 2$ dans laquelle est inclus un algorithme de génération de bordures pour un nombre d'articles N fixé (cf. algo1 p.10). L'itération se poursuit tant que N est trop petit pour permettre une construction complète sur le dernier niveau où la distribution donnée $\mathcal{S}^-$ est non nulle.

Algorithme 1 : GenerationBordures

```
Données : α ; S⁻
Sorties : D⁻, D⁺ ; BD⁻, BD⁺, N
1 N ← 1 ;
2 répéter
3 |   N ← N + 1 ;
4 |   (D⁻,D⁺,BD⁻,BD⁺, NtropPetit) ←
  |   ConstructionTousNiveaux(N,α,S⁻);
5 jusqu'à NON (NtropPetit) ;
```

Pour un nombre d'articles N donné, la construction de la bordure négative s'effectue niveau après niveau, dans l'ordre colex inverse, en partant du niveau le plus bas (composé des motifs comportant le moins d'articles) sur lequel la distribution de la bordure négative est non nulle. Lorsqu'on traite un niveau, tous les éléments de la bordure négative des niveaux inférieurs sont déjà choisis. La bordure positive se construit par ajout successif au rythme de l'intégration de nouveaux motifs dans la bordure négative.

La stratégie utilisée pour la construction de la bordure négative d'un niveau donné est l'intégration de grappes de motifs consécutifs dans l'ordre colex séparées par une juxtaposition de « trous de fréquents » bien choisis. La largeur des trous dépend du paramètre α donné. Plus α est grand, plus les trous sont larges. Cette stratégie permet d'assurer que **les motifs d'un niveau k insérés dans la bordure négative induisent des éléments dans la bordure positive qui sont $(\alpha - 2)$ niveaux au dessus**. Le résultat attendu est l'obtention de bordures négative et positive de distributions « éloignées » l'une de l'autre, l'éloignement étant d'autant plus grand que le paramètre α est grand.

Le contrôle de l'éloignement a une contre-partie : les motifs sont insérés dans la bordure négative par paquets insécables. Ceci explique les éventuels (petits) écarts entre la distribution de bordure négative construite $\mathcal{D}^-$ et la distribution $\mathcal{S}^-$ donnée en entrée.

Cas $\alpha = 1$ Il correspond à une construction avec insertion de blocs entiers consécutifs dans la bordure négative, sans trous de fréquents, du type de celle de [DDM05]. Du fait de l'insertion de blocs,

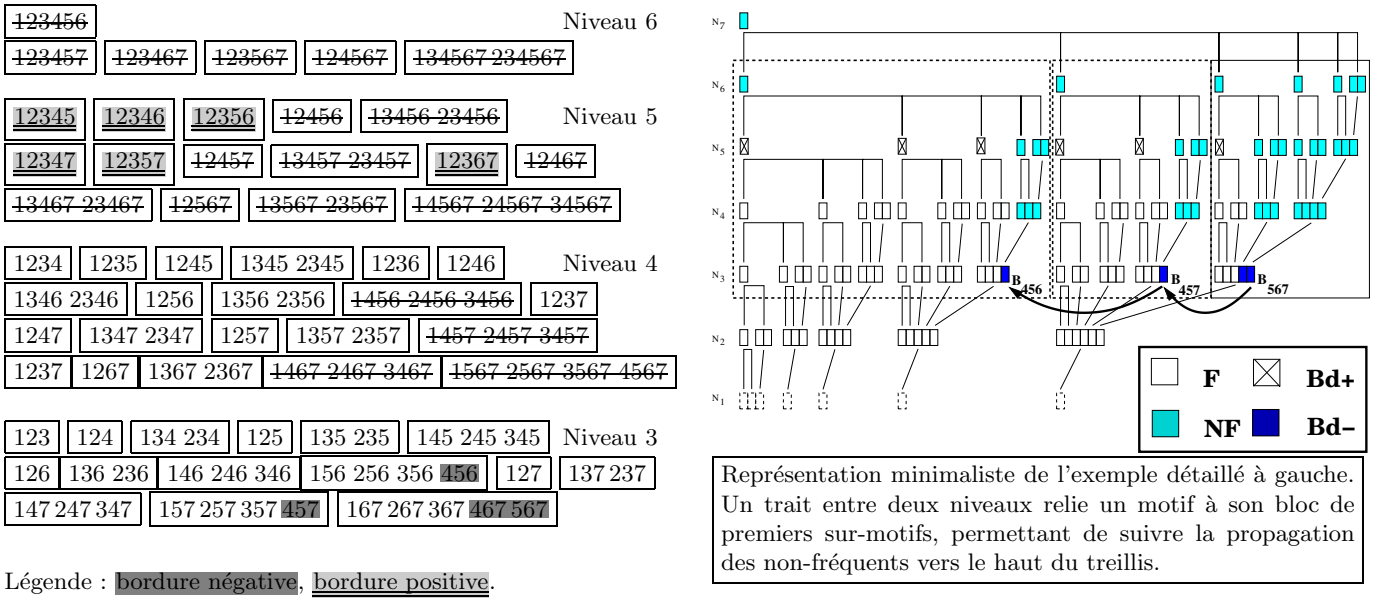

FIG. 4 – Exemple de construction d'une bordure négative sur un seul niveau avec un paramètre d'éloignement $\alpha = 4$ (algorithme 3 avec : $\alpha = 4$, $\mathcal{S}^- =< 0, 0, 4 >$, $N = 7$). Seuls les motifs de rang ≥ 4 dans leur bloc sont intégrés. La bordure positive correspondante est éloignée de $\alpha - 2 = 2$ niveaux.

la bordure générée est bien entendu différente de celle de [DDM05] car on approche la distribution donnée sans l'égaler.

Influence de α sur le nombre d'articles nécessaires Plus α est grand, plus les trous entre les plages de motifs de la bordure négative seront grands et plus le nombre d'articles N nécessaires à la construction sera grand. C'est là une des clés qui permet d'entrevoir la génération de jeux de données synthétiques réalistes.

5.2 Construction des motifs de la bordure négative du niveau k

L'algorithme 3 de la page 12 décrit complètement la stratégie utilisée pour la construction sur un niveau donné.

Le premier niveau traité est le niveau le plus bas pour lequel la distribution entrée en paramètre est non nulle (algo2 l.1). Le bloc de départ est le dernier bloc de ce niveau dans l'ordre colex (algo2 l.2). Pour un niveau k, on examine les blocs dans l'ordre colex inverse à partir d'un bloc déterminé à la fin de la construction pour le niveau précédent.

Seuls les blocs de taille supérieure ou égale à α sont « intégrés à la bordure négative ». Les blocs de taille strictement inférieure à α ne sont pas intégrés, ils sont laissés fréquents. L'intégration d'un bloc du niveau k consiste à ajouter à la bordure négative les motifs consécutifs du bloc ayant un rang supérieur ou égal à α (algo3 l. 5, 21 et 23). Il est alors simple de calculer la plage de motifs du niveau $k + \alpha - 2$ à mettre dans la bordure positive. Ceux-ci ne dépendent que de l'indice de croissance colex du motif caractérisant le bloc intégré et des articles du motif. Le nombre de motifs à rajouter dans la bordure positive est 1 sauf lorsqu'on intègre à la bordure négative l'unique motif de rang supérieur à α d'un bloc de taille α. Pour des formules plus détaillées, on se reportera au tableau récapitulatif de figure 5 page 13. Pour les démonstrations, on se reportera à [Ber07].

Exemple 8 *La figure 4 illustre ce procédé dans un cas simple où la distribution entrée est non nulle sur le seul niveau* 3 *et où le nombre d'articles est fixé à* 7*. Le paramètre d'éloignement* α *vaut* 4*, les motifs intégrés à la bordure négative sont donc tous de rang supérieurs à* 4 *dans leur bloc. Trois intégrations sont effectuées ; dans l'ordre,* $\mathcal{B}_{567}$*,* $\mathcal{B}_{457}$ *et enfin* $\mathcal{B}_{456}$*, ce qui donne* $\mathcal{B}d^- = \{567, 467, 457, 456\}$*.*

Deux représentations de cet exemple sont données : à gauche de la figure la représentation habituelle où les motifs sont énumérés et à droite la représenta-

Algorithme 2 : ConstructionTousNiveaux

Données : N ; α ; $\mathcal{S}^-$
Sorties : $\mathcal{D}^-$, $\mathcal{D}^+$; $\mathcal{BD}^-$, $\mathcal{BD}^+$; NtropPetit
/* pré-conditions : */
/* ${\mathcal{S}^-}_1 = 0$, N − plusPetitNiveau$\mathcal{S}^-$ $+1 \geq \alpha$ */
/* Construction d'une bordure négative de distribution proche de $\mathcal{S}^-$ et de la bordure positive "distante" de $\alpha - 2$ niveaux correspondante */

// Les initialisations
1 $\mathcal{D}^- \leftarrow < 0, \cdots, 0 >$; $\mathcal{D}^+ \leftarrow < 0, \cdots, 0 >$;
2 k ← plusPetitNiveau$\mathcal{S}^-$; X ← dernierColex(k,N) ;
3 NtropPetit ← **VRAI** ;
4 **répéter**
  /* Construction de la bordure négative pour le niveau k et de la bordure positive correspondante */
  /* Y est le bloc qui aurait été intégré dans ${\mathcal{BD}^-}_k$ si le traitement du niveau k s'était poursuivi. */
5 $({\mathcal{D}^-}_k, {\mathcal{D}^+}_{k+\alpha-2}, {\mathcal{BD}^-}_k, {\mathcal{BD}^+}_{k+\alpha-2},$ Y, existeY, NiveauComplet) ← ... ConstructionNiveau(k,X,${\mathcal{S}^-}_k$,α);
6 **si** existeY **ET** tailleBloc(Y) $= \alpha$ **alors**
    /* Changement pathologique de niveau */
7 W ← blocTaillePlusUnSuccesseur(Y) ;
8 d^+ ← W.indiceCC − Y.indiceCC $+1$;
9 ${\mathcal{D}^+}_{k+\alpha-1} \leftarrow {\mathcal{D}^+}_{k+\alpha-1} + d^+$;
10 (A, B) ← bornesBDPBriqueChangementNiveau (Y,W);
11 integreIntervalleABordure($\mathcal{BD}^+$,A,B) ;
    /* remplacement de Y */
12 (Y,existeY) ← blocTailleSuperieurePredecesseur(W) ;
13 **fin**
14 **si** existeY **alors**
15 X ← blocPremierSurMotifDernierElement(Y) ;
16 k ← k $+1$;
17 **fin**
18 **jusqu'à** k = plusGrandNiveau$\mathcal{S}^-$ $+1$ **OU NON** existeY ;
19 **si** (k = plusGrandNiveau$\mathcal{S}^-$ $+1$) **alors**
20 Finition par rajout de motifs dans $\mathcal{BD}^+$ uniquement, sur des niveaux supérieurs à plusGrandNiveau$\mathcal{S}^-$ $+ \alpha - 2$;
21 NtropPetit ← **FAUX** ;
22 **fin**
23 **si** k = plusGrandNiveau$\mathcal{S}^-$ **ET** NiveauComplet **alors**
24 NtropPetit ← **FAUX** ;
25 **fin**

Algorithme 3 : ConstructionNiveau

Entrées : k ; X ; α ; ${S^-}_k$
Sorties : ${\mathcal{D}^-}_k$; ${\mathcal{D}^+}_{k+\alpha-2}$; ${\mathcal{BD}^-}_k$; ${\mathcal{BD}^+}_{k+\alpha-2}$; Y ; existeY ; IntegrationBlocOK ;
/* Pré-conditions : */
/* tailleBloc(X) $\geq \alpha$, k = X.longueur */
/* Pour un niveau donné, construction de ${\mathcal{BD}^-}_k$ et de ${\mathcal{BD}^+}_{k+\alpha-2}$ correspondante. */
/* Le bloc $\mathcal{B}_X$ est le premier intégré à ${\mathcal{BD}^-}_k$ puis le bloc courant */
/* Le bloc $\mathcal{B}_Y$ est le bloc suivant $\mathcal{B}_X$ dans le traitement */

1 NiveauComplet ← **FAUX** ;
2 existeY ← **FAUX** ;
3 **répéter**
4 p ← tailleBloc(X) ;
  /* Gain potentiel sur bordure négative */
5 d^- ← p $- \alpha + 1$;
  /* Dépassement sur niveau k ? */
6 **si** ${\mathcal{D}^-}_k + d^- \geq {S^-}_k$ **alors**
7 NiveauComplet ← **VRAI**
8 **fin**
  /* Intégration du bloc et calcul du bloc suivant à traiter */
9 **si** p ≥ 2 **alors**
10 **si** α = p **alors**
11 d^+ ← X.indiceCC ;
12 (Y,existeY) ← blocTailleSuperieurePred (X) ;
13 **sinon**
14 d^+ ← 1 ;
15 (Y,existeY) ← blocPredecesseur (X) ;
16 **fin**
17 **sinon**
18 d^+ ← X.indiceCC ;
19 (Y,existeY) ← blocPredecesseur (X) ;
20 **fin**
  /* Mise à jour des distributions de bordures */
21 ${\mathcal{D}^-}_k \leftarrow {\mathcal{D}^-}_k + d^-$;
22 ${\mathcal{D}^+}_{k+\alpha-2} \leftarrow {\mathcal{D}^+}_{k+\alpha-2} + d^+$;
  /* met les motifs de rang $\geq$ à α du bloc $\mathcal{B}_X$ dans $\mathcal{BD}^-$ */
23 integreFinBlocABordure($\mathcal{BD}^-$,X,α) ;
  /* met un ou plusieurs motifs du niveau k + $\alpha - 2$ dans $\mathcal{BD}^+$ */
24 (A, B) ← extremitesBordurePositiveBrique (X,α) ;
25 integreIntervalleABordure($\mathcal{BD}^+$,A,B);
26 **si** existeY **alors** X ← Y ;
27 **jusqu'à** NiveauComplet **OU NON** existeY ;

J est l'indice de croissance colex de $p(p+1)x_3\cdots x_k$. Il est toujours supérieur ou égal à 2.

Pour un bloc de taille $p=1$

α	*niv.*	$\sharp$	motifs à mettre dans $\mathcal{B}d^+$
1	$k-1$	J	$\geq 1\cdots(J-1)x_{J+1}\cdots x_k$ $\leq 2x_3\cdots x_k$

Pour un bloc de taille $p=2$

α	*niv.*	$\sharp$	motifs à mettre dans $\mathcal{B}d^+$
1	$k-1$	1	$3x_3\cdots x_k$
$p=2$	k	J	$\geq 12\cdots Jx_{J+1}\cdots x_k$ $\leq 13x_3\cdots x_k$

C'est un cas particulier du cas $p\geq 3$ exposé ci-dessous.

Pour un bloc de taille $p\geq 3$

α	*niv.*	$\sharp$	motif(s) à mettre dans $\mathcal{B}d^+$
1	$k-1$	1	$(p+1)x_3\cdots x_k$
2	k	1	$1(p+1)x_3\cdots x_k$
3	$k+1$	1	$12(p+1)x_3\cdots x_k$
$\vdots$			
α	$k+\alpha-2$	1	$1\cdots(\alpha-1)(p+1)x_3\cdots x_k$
$\vdots$			
$p-1$	$k+p-3$	1	$1\cdots(p-2)(p+1)x_3\cdots x_k$
p	$k+p-2$	J	$\geq 12\cdots(J+p-2)x_{J+1}\cdots x_k$ $\leq 1\cdots(p-1)(p+1)x_3\cdots x_k$

FIG. 5 – **Récapitulatif :** Conséquence sur $\mathcal{B}d^+$ de l'ajout dans $\mathcal{B}d^-$ des motifs de $\mathcal{B}_{p(p+1)x_3\cdots x_k}$ à partir du rang α : ajout de $\sharp$ motifs sur le niveau niv.

tion minimaliste où les motifs sont symbolisés par un rectangle et l'application Θ *« a pour bloc de premiers sur-motifs » par un trait partant d'un niveau vers le niveau du dessus.*

Il est simple sur ce petit exemple de trouver les non fréquents induits dans les niveaux supérieurs par les deux motifs non fréquents du bloc $\mathcal{B}_{567}$ *puis par celui de* $\mathcal{B}_{457}$ *et enfin par celui de* $\mathcal{B}_{456}$*. Il sera utile de se convaincre sur cet exemple de l'intérêt de la proposition 4. En effet, l'itération qui intègre le bloc* $\mathcal{B}_{567}$ *consiste en fait à fixer le statut fréquent ou non fréquent de tous les motifs contenus dans le grand rectangle de droite. La proposition 4 assure que l'itération suivante qui fixe le statut fréquent ou non fréquent de tous les motifs contenus dans le grand rectangle du milieu ne modifie pas les statuts fixés à l'itération précédente sur le rectangle de droite (par exemple, le motif de rang 4 de* $\mathcal{B}_{457}$ *n'induit aucun motif non fréquent supplémentaire*

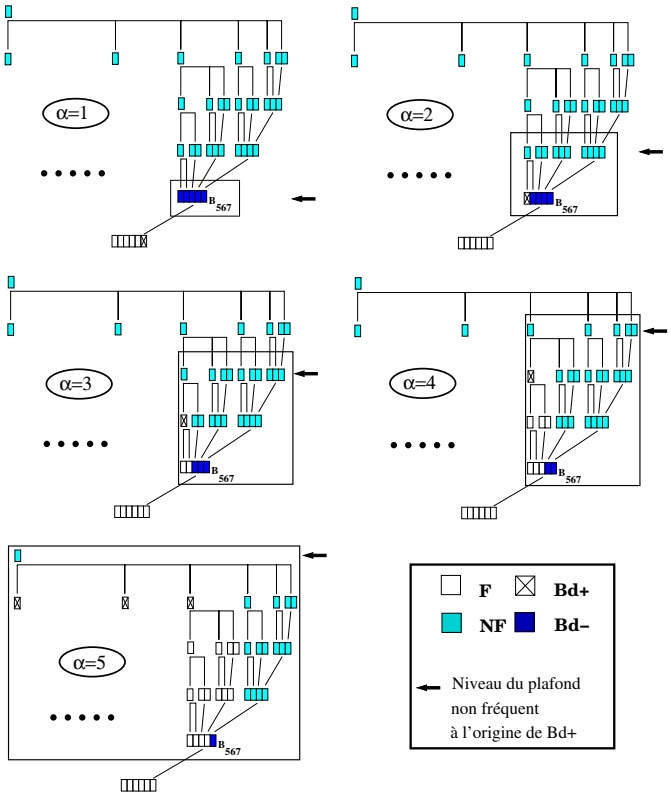

FIG. 6 – Conséquence sur $\mathcal{B}d^+$ de l'ajout dans $\mathcal{B}d^-$ des motifs de $\mathcal{B}_{p(p+1)x_3\cdots x_k}=\mathcal{B}_{567}$ à partir du rang α pour α variant de 1 à $p=5$. Le nombre de motifs consécutifs (dans l'ordre colex) insérés dans $\mathcal{B}d^-$ détermine le niveau auquel le(s) motif(s) de $\mathcal{B}d^+$ induit(s) appartien(nen)t

parmi les motifs de la « brique » à sa droite).
Les motifs de la bordure positive apparaissent $\alpha-2=2$ *niveaux au dessus et la bordure positive* $\mathcal{B}d^+$ *est l'ensemble* $\{12367, 12357, 12347, 12356, 12346, 12345\}$. ♢

Le fait d'intégrer dans la bordure négative uniquement des motifs de blocs de taille supérieure à α et de laisser de côté tous les blocs de taille strictement inférieure à α suppose de connaitre la façon dont les blocs s'enchaînent. On indique ci-dessous comment on passe de l'intégration d'un bloc dans la bordure négative à l'intégration suivante.

Bloc $\mathcal{B}_Y$ traité après le dernier bloc $\mathcal{B}_X$ intégré **Pour un bloc $\mathcal{B}_X$ de taille supérieure ou égale à 2** Un bloc de taille $p\geq 2$ est toujours

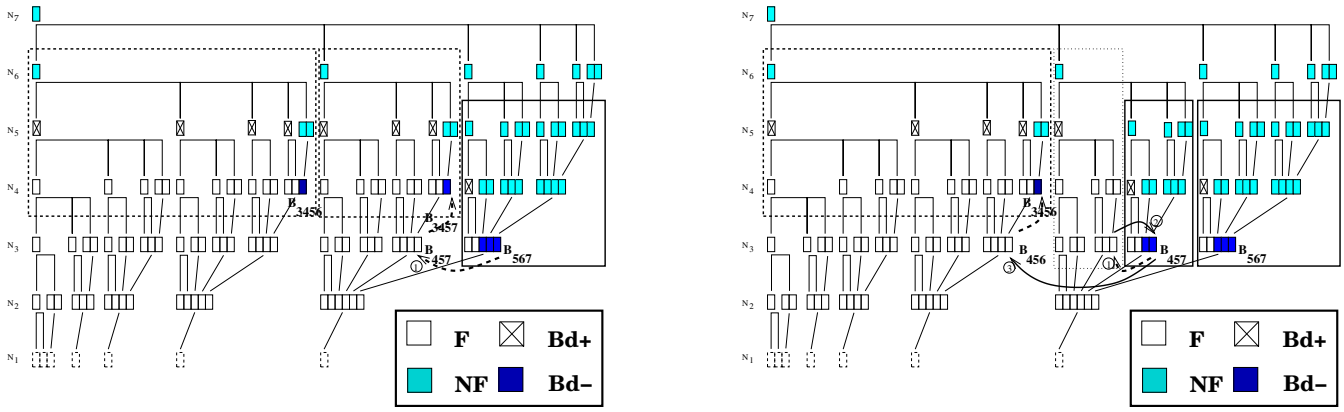

FIG. 7 – Changement de niveau (de k à $k+1$). Ici $\alpha = 3$, $N = 7$. Le traitement est différent selon que la taille du bloc $\mathcal{B}_Y$ qui aurait dû être traité au niveau k est $\geq \alpha$ (à gauche) ou $= \alpha$ (à droite).

suivi dans l'ordre colex inverse d'un bloc de taille $p-1$. Donc tant qu'un bloc a une taille supérieure ou égale à $\alpha+1$, le bloc suivant à intégrer est le bloc qui le suit dans l'ordre colex inverse (algo3 l.15). Si le bloc a une taille égale à α, alors le bloc suivant dans l'ordre colex inverse n'est pas intégrable, il faut trouver le premier bloc de taille supérieure ou égale à α qui convient (algo3 l.12). La proposition 5 page 9 donne la formule permettant de déterminer ce bloc sans examiner tous les blocs intermédiaires de taille inférieure à α.

Pour un bloc $\mathcal{B}_X$ de taille égale à 1 Un bloc de taille $p = 1$ n'est examiné que lorsque $\alpha = 1$. Dans le cas $\alpha = 1$, tous les blocs sont intégrables. Le bloc suivant dans le traitement est le bloc suivant dans l'ordre colex inverse (algo3 l.19).

Condition d'arrêt de construction sur le niveau k

L'intégration d'un bloc de taille p apporte $p - \alpha + 1$ éléments à la bordure négative. On passe au niveau suivant lorsque le nombre de motifs d_k^- du niveau k insérés dans la bordure négative dépasse le cardinal s_k^- du niveau k de la distribution $\mathcal{S}^-$ donnée. On « approche » ainsi $\mathcal{S}^-$ par valeur supérieure, au mieux que le permet la stratégie de construction utilisée.

5.3 Passage au niveau supérieur $k+1$

Lorsque les motifs du niveau k de la bordure négative sont tous déterminés, on connait le bloc $\mathcal{B}_Y$ qui aurait dû être traité si l'intégration des motifs sur le niveau k s'était poursuivie. Le premier bloc du niveau $k+1$ qu'il est possible d'intégrer à la bordure négative dépend de la taille de ce bloc $\mathcal{B}_Y$. Deux cas sont possibles. Le cas le plus simple est celui où la taille de $\mathcal{B}_Y$ est supérieure ou égale à $\alpha+1$. Dans le cas où la taille de $\mathcal{B}_Y$ est égal à α, le traitement se fait en deux temps en rajoutant dans un premier temps des motifs dans la bordure positive pour se ramener à une sitation où un nouveau bloc $\mathcal{B}_Y$ est de taille $\alpha+1$.

Exemple 9 *La figure 7 présente un exemple pour chacun de ces deux cas. Le paramètre α vaut 3.*

• *À gauche. Le dernier bloc intégré est le bloc $\mathcal{B}_{567}$ du niveau 3. Ce bloc est de taille $5 > \alpha$, le bloc $\mathcal{B}_Y$ suivant dans le traitement si on poursuivait sur ce niveau 3 serait donc le bloc prédécesseur $\mathcal{B}_{457}$ qui est de taille 4. Le premier bloc intégré sur le niveau au dessus est le bloc de premier sur-motif de 457, i.e. $\mathcal{B}_{3457}$. On poursuit ensuite sur le niveau 4 de la façon préconisée dans la section 5.2. On a* $\mathcal{B}d^- = \{567, 467, 367, 3457, 3456\}$ *et* $\mathcal{B}d^+ = \{1267, 12457, 12357, 12347, 12456, 12356, 12346, 12345\}$.

• *À droite. Le dernier bloc intégré est le bloc $\mathcal{B}_{457}$ du niveau 3. Ce bloc est de taille $4 > \alpha$, le bloc $\mathcal{B}_Y$ suivant dans le traitement est donc le bloc prédécesseur $\mathcal{B}_{347}$ qui est de taille 3. On se ramène au premier cas en remplaçant $\mathcal{B}_{347}$ par le bloc de taille $\geq (\alpha+1)$ qui le précède, à savoir par le $\mathcal{B}_{456}$ de taille 4. On poursuit alors comme dans le premier cas sur le niveau au dessus en intégrant le bloc*

$\mathcal{B}_{3456}$ *de premiers sur-motifs du motif* 456. *Notons que la taille de ce bloc est* $3 = \alpha$. *Il y a un manque entre les deux blocs intégrés que l'on comble en laissant se propager les non-fréquents sans ajouter de motifs à la bordure négative. Ceci amène un motif supplémentaire dans la bordure positive. On a :*
$\mathcal{B}d^- = \{567, 467, 367, 457, 357, 3456\}$ *et*
$\mathcal{B}d^+ = \{1267, 1257, 12347, 12456, 12356, 12346, 12345\}$. ♢

Cas où le bloc $\mathcal{B}_Y$ est de taille supérieure ou égale à $\alpha + 1$ (fig.7 page 14 à gauche).

Si le bloc $\mathcal{B}_Y$ est de taille supérieure ou égale à $\alpha+1$, alors le bloc de premiers sur-motifs du dernier motif Y de $\mathcal{B}_Y$ appartient au niveau $k+1$ et sa taille est supérieure ou égale à α. Ce bloc est le bloc de départ pour l'intégration dans la bordure négative du niveau $k + 1$ (algo2 l.15).

Ce cas de figure se rencontre lorsque le dernier bloc intégré du niveau k est de taille supérieure ou égale à $\alpha + 2$ mais aussi lorsque le dernier bloc intégré du niveau k est de taille α. En effet, sur un niveau donné, le bloc suivant dans le traitement un bloc de taille $\alpha \geq 2$ est le premier bloc de taille supérieure ou égale à α ; cette taille peut donc être supérieure ou égale à $\alpha + 1$. Par ailleurs un bloc de taille 1 peut être suivi dans l'ordre colex inverse d'un bloc de n'importe quelle taille.

Cas où le bloc $\mathcal{B}_Y$ est de taille α (fig. 7 page 14 à droite)

(a) Premier bloc traité sur le niveau $k+1$ Si le bloc $\mathcal{B}_Y$ est de taille égale à α et que $\alpha \geq 2$, alors le bloc de premiers sur-motifs du dernier motif de $\mathcal{B}_Y$ appartient au niveau $k+1$ et sa taille est égale à $\alpha-1$. Ce bloc n'est pas intégrable. Si le bloc $\mathcal{B}_Y$ est de taille α et que α vaut 1, le bloc de premier sur-motif de $\mathcal{B}_Y$ est de taille 1 mais il est non fréquent en tant que bloc de premier sur-motif du bloc de taille 2 qui a précédé dans le traitement en ordre colex inverse. Par conséquent, ce bloc n'est pas intégrable. Dans les deux cas, la stratégie employée lorsque $\mathcal{B}_Y$ est de taille supérieure ou égale à $\alpha+1$ n'est pas utilisable.

Dans le cas où $\mathcal{B}_Y$ est de taille α, on détermine le bloc $\mathcal{B}_Z$ de taille supérieure ou égale à $\alpha+1$ qui suit $\mathcal{B}_Y$ dans l'ordre colex inverse. La proposition 6 page 9 donne la formule permettant de déterminer ce bloc sans examiner tous les blocs intermédiaires de taille inférieure à $\alpha+1$. Comme dans la proposition 6, on détermine d'abord le premier bloc $\mathcal{B}_W$ de taille $\alpha+1$ qui suit le bloc $\mathcal{B}_Y$ dans l'ordre colex ainsi que l'indice de croissance colex de W. Le bloc Z est alors le premier bloc de taille supérieure ou égale à $\alpha + 1$ qui précède le bloc $\mathcal{B}_W$ (algo2 l.7 puis l.12).

Le bloc de premiers sur-motifs du dernier motif de Z appartient au niveau $k + 1$ et sa taille est supérieure ou égale à α. Ce bloc est le bloc de départ pour l'intégration dans la bordure négative du niveau $k + 1$ (algo2 l.15).

Ce cas de figure se rencontre lorsque le dernier bloc intégré à la bordure négative est de taille $\alpha + 1$ mais aussi lorsque le dernier bloc intégré est de taille α. En effet, lorsque le dernier bloc intégré est de taille α, le bloc suivant dans le traitement est le premier bloc de taille supérieure ou égale à α qui suit dans l'ordre colex inverse. La taille de ce bloc peut être exactement égale à α.

(b) Ajout de motifs dans la bordure positive Il faut bien entendu prendre en compte le fait que des blocs intermédiaires de taille inférieure ou égale à α ont été laissés de côté entre les blocs $\mathcal{B}_Z$ et $\mathcal{B}_Y$. Cela se fait très simplement en rajoutant des motifs du niveau $k+\alpha-1$ dans la bordure positive (algo2 l.7 à l.11).

Notons $\mathcal{B}_{W^1}$ le bloc de taille 1 qui suit $\mathcal{B}_Z$ dans l'ordre colex. Le bloc $\mathcal{B}_Y$ est compris entre les blocs $\mathcal{B}_{W^1}$ et $\mathcal{B}_W$. Au niveau $k+\alpha-1$, le bloc $\mathcal{B}_Y$ de taille α génère par itération sur les niveaux un seul bloc

Y : motif de rang p dans un bloc de taille p du niveau k, J indice de croissance colex de Y, i.e.
$Y = p(p+1)\cdots(p+J-1)y_{J+1}\cdots y_k$ avec $y_{J+1} > p+J$.
$W := (p+1)(p+2)\cdots(p+J)y_{J+1}\cdots y_k$
H l'indice de croissance colex de W.
Afin de pouvoir traiter un bloc de taille $\alpha+1$, on installe une « brique » dont la base est composé des motifs de $[1\cdots Hy_{H+1}\cdots y_k, p(p+1)\cdots(p+J-1)y_{J+1}\cdots y_k]$ qui sont tous fréquents. La conséquence sur $\mathcal{B}d^+$:

niv.	♯	Motifs à mettre dans $\mathcal{B}d^+$
$k+p-1$	$H-J+1$	$\geq 1\cdots(H+p-1)y_{H+1}\cdots y_k$ $\leq 1\cdots(J+p-1)y_{J+1}\cdots y_k$

FIG. 8 – **Récapitulatif** pour le changement de niveau. Cas où la taille du bloc $\mathcal{B}_Y$ est $p = \alpha$.

de premiers sur-motifs et il est de taille 1. Notons le $\Theta^{(\alpha-1)}(\mathcal{B}_Y)$. Le bloc lui aussi génère par itération sur les niveaux un seul bloc de premiers sur-motifs et il est de taille 1. Notons le $\Theta^{(\alpha-1)}(\mathcal{B}_{W^1})$. Les motifs à inclure dans la bordure positive sont les motifs des blocs de taille 1 du niveau $k+\alpha-1$ qui vont de $\Theta^{(\alpha-1)}(\mathcal{B}_{W^1})$ à $\Theta^{(\alpha-1)}(\mathcal{B}_Y)$. Les formules des deux extrémités de l'intervalle à inclure sont données dans la figure 8. On voit que le nombre de motifs compris entre ces deux bornes dépend seulement des indices de croissance colex de Y et de W.

5.4 Finalisation de la bordure positive

Du fait de l'existence de la boucle sur N (algo1), il existe un plus petit nombre d'articles N pour lequel, sur tous les niveaux, la distribution de la bordure négative calculée finit par être supérieure ou égale à la distribution entrée y compris sur le dernier niveau non nul de $\mathcal{S}^-$ ($d^-_{kmax} \geq s^-_{kmax}$). On a alors trouvé, pour le α fixé en entrée, le nombre d'articles N le plus petit nécessaire à la construction d'une bordure négative de distribution proche de la distribution donnée en entrée.

Il se peut que le dernier bloc intégré à la bordure négative amène l'intégration dans la bordure positive d'éléments comprenant le premier motif du niveau $kmax+\alpha-2$. C'est le cas où l'on n'aurait de tout façon pas pu poursuivre la construction même si s^-_{kmax} n'avait pas été atteint car il n'existe pas de bloc $\mathcal{B}_Y$ successeur de $\mathcal{B}_X$ dans le traitement. La construction est terminée.

Dans le cas contraire, on aurait pu poursuivre la construction si s^-_{kmax} avait été plus grand. Il reste alors des motifs dont il faut fixer le statut fréquent ou non fréquent. Le choix a été fait de prendre fréquents les motifs compris dans l'ordre colex entre le premier motif du niveau $kmax$ et le motif Y. De ce fait, l'algorithme se termine par le calcul de la bordure positive des motifs compris dans l'ordre colex entre le premier motif du niveau $kmax$ et le motif Y. Ce calcul, qui utilise le même genre de formules que celles utilisées pour les changements de niveau, n'est pas détaillé dans ce document.

5.5 Validité de l'algorithme

Sur les exemples présentés dans les figures 4 et 7, on voit que chaque intégration de bloc dans la bordure négative revient à fixer le statut fréquent ou non fréquent d'un certain nombre de motifs qu'on a entouré d'un grand rectangle. Cet ensemble de motifs forme une « brique de construction ». Trois types de briques ont été mis en évidence. La « *brique de niveau normale de paramètre α* » a sa base constituée d'un bloc $\mathcal{B}_X$ de taille strictement supérieure au paramètre α. Cette brique normale a une hauteur de α niveaux et est constituée de tous les motifs des ensembles $\Theta^\gamma(\mathcal{B}_X)$ pour γ de 0 à $\alpha-1$ (cf. définition 11 de Θ page 8). Le nombre de motifs intégrés à la bordure négative est supérieur ou égal à 2 ce qui induit un seul motif supplémentaire dans la bordure positive. La « *brique de niveau élargie de paramètre α* » a sa base constituée d'un bloc de taille égale à α qu'on intègre dans la bordure négative et aussi de tous les blocs précédents consécutifs et de taille strictement inférieure à celle du bloc qu'on intègre. Si l'on note $\mathcal{B}$ l'ensemble de tous ces motifs, cette brique élargie a une hauteur de α niveaux et est constituée de tous les motifs des ensembles $\Theta^\gamma(\mathcal{B})$ pour γ de 0 à $\alpha-1$. Le nombre de motifs intégrés à la bordure négative est 1, ce qui induit non pas un mais plusieurs motifs supplémentaires dans la bordure positive. Enfin la « *brique de changement de niveau de paramètre α* » permet de générer une bordure positive un niveau au dessus de celui atteint à l'intégration précédente sans ajout de motif dans la bordure négative. Elle a sa base constituée d'un bloc de taille α et aussi de tous les blocs de taille inférieure ou égale qui le précèdent.

Proposition 7 *Sur un niveau k donné, les briques de niveau normales et élargies de paramètres α sont juxtaposables.* ■

En effet, les briques ne se superposent pas (conséquence de la proposition 3 page 8). Par ailleurs, une brique de niveau de paramètre α crée à sa droite sur le niveau du dessus des motifs non fréquents de rang supérieur à α (proposition 4 page 9) donc on ne modifie pas les briques déjà installées à droite. Un exposé plus complet des « briques » et des démonstrations associées se trouve dans [Ber07]. Ceci amène à donner le théorème suivant :

Théoreme 1 *Étant donnés une distribution de bor-*

dure négative $\mathcal{S}^-$ et un entier α, l'algorithme 1 génère une bordure positive $\mathcal{B}d^+$ et une bordure négative $\mathcal{B}d^-$ de distribution respective $\mathcal{D}^+$ et $\mathcal{D}^-$ telles que :
(a) $\mathcal{D}^-$ « s'approche au plus prêt » de $\mathcal{S}^-$. Une majoration large donne $d_k^- - s_k^- \leq (N-k+1)-(\alpha-1)$.
(b) $\mathcal{B}d^+$ est constituée de k-motifs « provenant » de l-motifs de la bordure négative pour l tel que $l \leq k - \alpha + 2$. ■

La preuve du théorème est omise. C'est une conséquence des propositions données dans les sections précédentes [Ber07].

Remarque 3 En réalité, dans la mesure où la taille du bloc qu'on intègre le permet, on peut faire suivre dans l'ordre colex inverse une brique de niveau de paramètre α par une brique de niveau de paramètre supérieur à α. Par ailleurs, si α est le paramètre de la dernière brique du niveau k, le paramètre de la première brique du niveau $k+1$ peut être $\alpha - 1$. L'algorithme exposé dans cet article est en fait un cas particulier de juxtaposition de briques de paramètre α constant.

Complexité de l'algorithme 2. Sur un niveau k donné, le calcul du bloc prédécesseur, le calcul du bloc prédécesseur de taille supérieure ou égale, l'ajout d'un motif dans la bordure négative, l'ajout d'un motif dans la bordure positive, le calcul permettant de trouver le premier motif à traiter sur le niveau suivant sont des opérations en $\mathcal{O}(k)$. La construction de la bordure négative sur un niveau se fait au pire en d_k^- itérations. Si l'on note respectivement k_{min} et k_{max} le plus petit et le plus grand niveau sur lequel la distribution entrée $\mathcal{S}^-$ est non nulle, on a :
$\sum_{k=k_{min}}^{k_{max}} k = (k_{min} + k_{max})\ (k_{max} - k_{min} + 1)/2$.
L'algorithme 2 est finalement en $\mathcal{O}(k_{max}^2 \max_k(d_k^-))$.

6 Implémentation et expérimentations

6.1 Implémentation

L'algorithme a été développé en C++ en utilisant la *STL*. L'implémentation ne nécessite pas de structures de données élaborées, en particulier pour décrire les blocs. L'algorithme passe simplement d'un motif caractérisant un bloc (celui de rang le plus grand dans le bloc) à un motif caractérisant un autre bloc par des calculs utilisant les articles du motif et reprenant les formules données dans les propositions 5, 6 et dans les récapitulatifs des figures 5, 8. Par ailleurs, l'armature du treillis en ordre colex constituée par la relation « a pour bloc de premiers sur-motifs » n'apparait pas explicitement dans l'implémentation. Elle est cachée dans les calculs faits pour passer d'un motif à l'autre.

On notera que les bordures obtenues expérimentalement par l'implémentation ont toutes été validées grâce à la propriété de dualisation présentée dans [MT97].

6.2 Expérimentations

Pour les problèmes d'extraction de motifs fréquents, douze jeux de données issus d'applications réelles sont communément utilisés dans les bancs d'essais [FIMIRep]. Comme l'a montré l'étude expérimentale réalisée dans [FMP05], ces jeux de données peuvent être classés en trois types en fonction de la distribution de leurs bordures. Les expérimentations réalisées ont eu pour objectif de générer des jeux de données respectant ces différents types, et ayant donc des caractéristiques proches des données réelles.

Protocole expérimental Chacun des jeux de données a été étudié pour plusieurs seuils de support représentatifs (en % sur les figures). Puis, pour chacun d'entre eux, la distribution de la bordure négative et différentes valeurs de α ont été utilisées comme paramètres d'entrée de notre algorithme. Dans un second temps, afin d'avoir des caractéristiques les plus proches possible des données réelles, α a été fixé à $d+2$, $d \geq -1$. Pour rappel, la valeur d correspond à la « distance » séparant les deux bordures, i.e. à la différence entre le niveau ayant le plus de motifs pour la bordure positive et celui ayant le plus de motifs pour la bordure négative. Notons que toutes ces informations sur les jeux de données réels sont disponibles en ligne sur [Fexp].

Dans la suite, nous présentons les résultats obtenus à partir de Chess et Connect. Ces jeux de données sont représentatifs des principaux types ob-

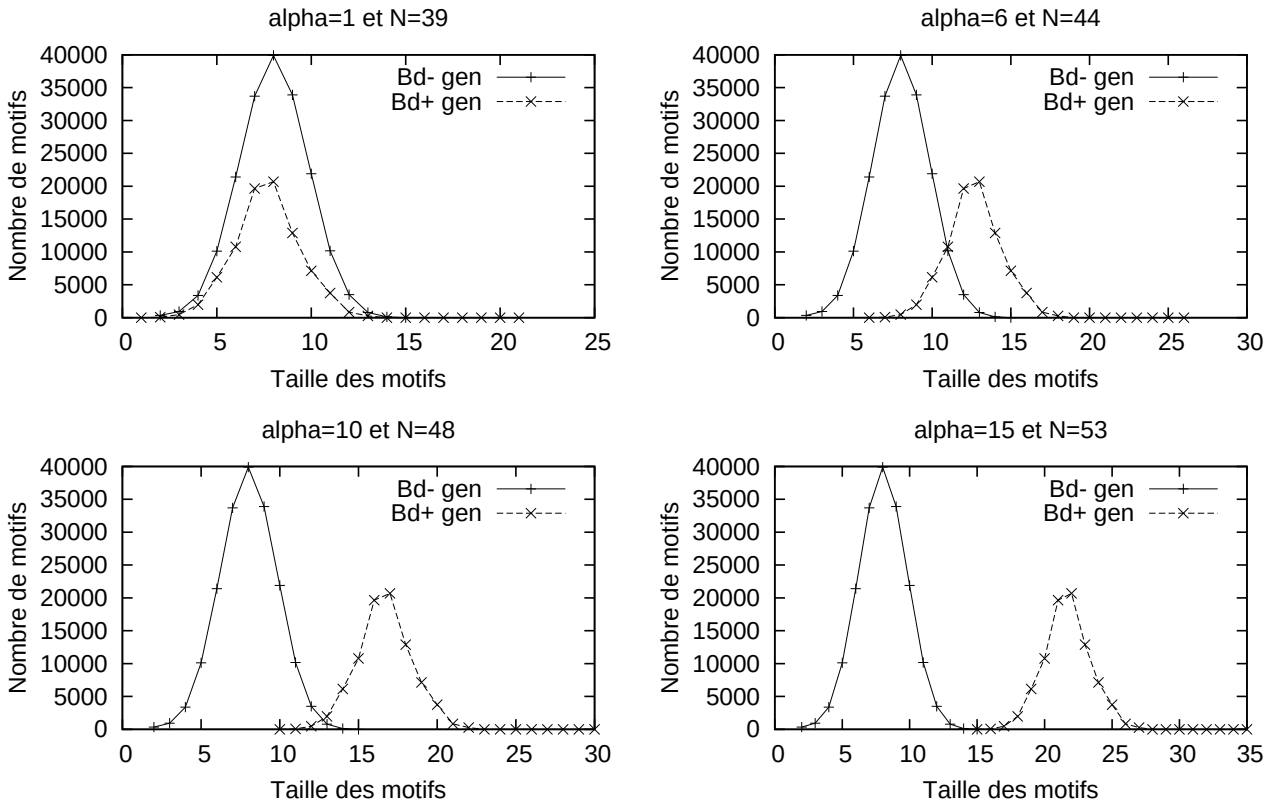

FIG. 9 – Influence de α pour Chess 30% ($\alpha = d + 2$)

servés dans [FMP05]. De plus, ils ont été utilisés dans [RMZ05] et [DDM05] pour montrer l'intérêt des approches proposées.

Les expérimentations ont été réalisées sur un Pentium IV 3.2 Ghz avec 1 Go de Ram. L'environnement utilisé était Windows XP, et le compilateur était gcc 3.4.2 (mingw). Notons que le temps d'exécution de l'algorithme a toujours été de l'ordre de la seconde lors de nos expérimentations.

Influence de alpha

La figure 9 montre l'influence de α passé en paramètre de l'algorithme de génération pour une distribution S^- en entrée identique à celle de Chess 30%. Comme attendu, la distribution de la bordure négative (notée « Bd- gen ») reste inchangée, et α permet de modifier la distance entre les deux bordures générées. Toutefois, même si la distance évolue, la courbe représentant la bordure positive (notée « Bd+ gen ») garde la même allure. De plus, il apparaît que lorsque α augmente, le nombre d'articles N utilisés pour construire les bases de données synthétiques augmente lui aussi, ce qui est normal puisque le nombre d'articles borne la taille des motifs de la bordure positive. D'après les expérimentations réalisées, cette augmentation semble linéaire.

Notons que lorsque $\alpha = 1$, la bordure positive est située un niveau en dessous de la bordure négative (figure 9 en haut à gauche). Ce cas correspond au type de jeux de données obtenus par les approches proposées dans [RMZ03, DDM05], mais n'est jamais observé dans les données réelles utilisées par la communauté [Fexp].

Comparaison avec les données réelles

Comme le montre la figure 10 page 19, les distributions de bordures négatives générées (notées « Bd- gen ») sont très proches des distributions données en entrée (notées « Bd- »). Les courbes paraissent même confondues dans la majorité des cas. La distance entre les deux bordures est elle aussi respectée. Par exemple, lorsque cette valeur est fixée à 4 pour Chess 60%, la distance entre les

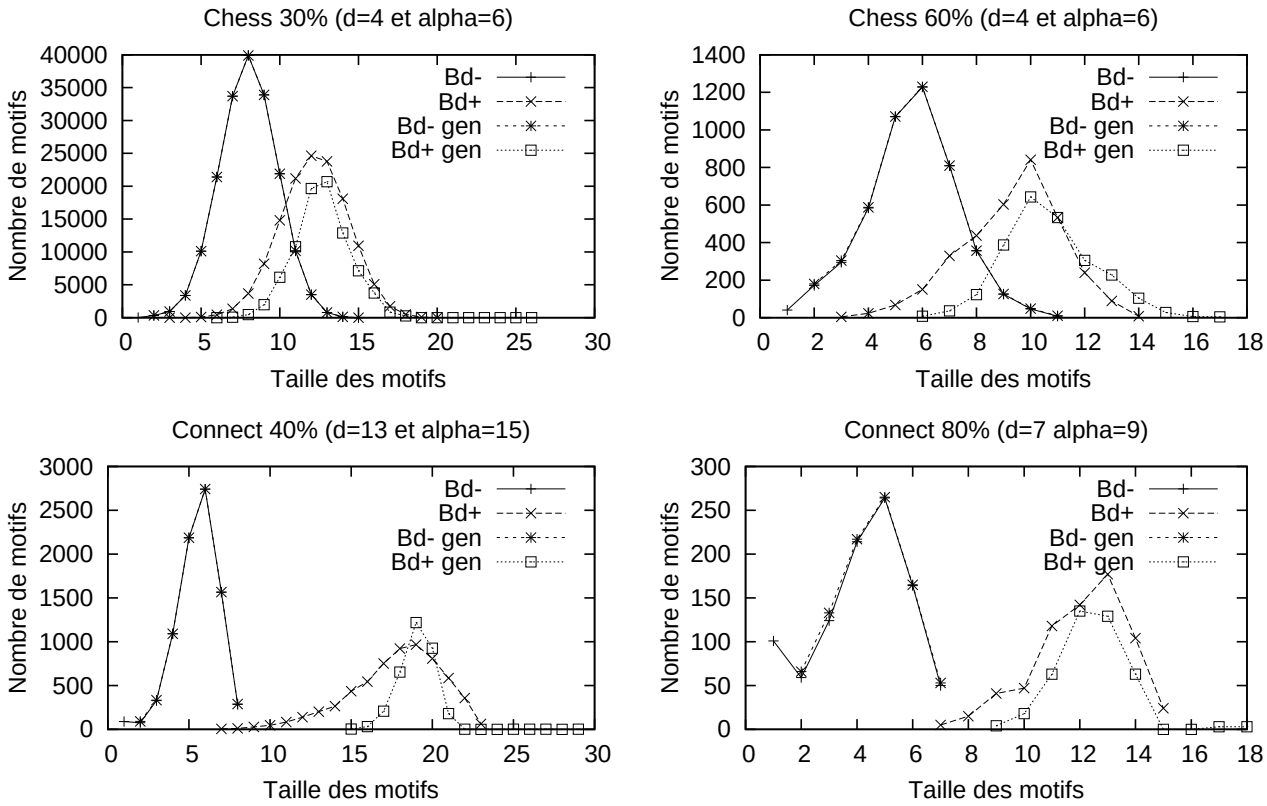

FIG. 10 – Exemples de bordures générées pour Chess et Connect

bordures générées est bien de 4 niveaux. De même, le jeu de données généré pour Connect 80% a bien une distance de 7 entre sa bordure négative et sa bordure positive.
Plus généralement, les distributions obtenues sont proches des distributions des jeux de données originaux. Les bordures obtenues à partir des caractéristiques de Chess 30% sont par exemple dans ce cas. La différence entre la bordure positive originale et celle générée peut être dans certains cas un peu plus importante, comme par exemple pour Connect 40%. Toutefois, les courbes conservent la même "forme". De plus, le type du jeu de données n'est pas remis en cause. En effet, la bordure positive générée pour Connect 40% est toujours sous la forme d'une courbe « en cloche », et la distance entre les deux bordures reste identique.

Le tableau 1 permet d'étudier le nombre d'articles fréquents (noté N) obtenus dans les jeux de données synthétiques construits. Seuls les articles fréquents sont considérés car les autres sont généralement ignorés au moment de la lecture de la base

	nb articles fréquents dans jdd réel	nb articles générés N
Chess 30%	50	44
Chess 60%	34	32
Connect 40%	41	39
Connect 80%	28	32

TAB. 1 – Tableau comparant le nombre d'articles entre jeux de données réels et synthétiques

de données. Comme le montre le tableau, le nombre d'articles utilisés pour générer ces jeux de données synthétiques est très proche du nombre d'articles fréquents des jeux de données originaux.

7 Conclusions et perspectives

Dans ce papier, nous avons proposé une nouvelle approche de génération de jeux de données synthétiques pour les problèmes équivalents aux motifs fréquents. Techniquement, cette contribution porte sur une structuration fine de l'ordre colex qui per-

met, par la notion de blocs, de « décaler » les bordures positive et négative. L'idée est de laisser se propager des éléments de la bordure positive dans des niveaux supérieurs, sans altérer la distribution sous-jacente de la bordure négative.

L'intérêt est de pouvoir reproduire synthétiquement les jeux de données réels, ce qui apporte une nouvelle solution pour la construction de jeux de données synthétiques. Cette contribution permet d'envisager pour la classe de problèmes d'énumération « représentables par des ensembles », des générations de jeux de données synthétiques permettant de développer de véritable campagnes de tests. À terme, il s'agit de mieux comprendre les raisons du succès ou de l'échec de tel ou tel algorithme.

Nous donnons ci-dessous quelques éléments de perspectives pour ce travail.

L'algorithme exposé dépend d'un paramètre α constant qui rend compte d'un éloignement entre les distributions de bordures positive et négative. Les tests expérimentaux ont été satisfaisants, même si l'on peut noter, par exemple figure 10 en haut à droite, que nous nous approchons certes des distributions réelles mais en restant assez loin du nombre réel d'éléments. Si l'on choisit de définir la « signature » d'un jeu de données par les deux distributions de bordures positive et négative, une bonne simulation d'un jeu de données quelconque doit forcément prendre en compte la distribution de bordure positive complète et pas seulement une notion peu précise d'éloignement. Dans cette optique, il semble intéressant de rendre variable le paramètre α de l'algorithme présenté. La remarque 3 page 17 invite à aller dans ce sens. L'étape suivante est donc de proposer un algorithme pour lequel l'entrée est constituée des deux distributions des bordures négative et positive. Le paramètre α ne sera plus une entrée mais sera ajusté à chaque intégration de blocs dans la bordure négative pour engendrer des motifs dans la bordure positive sur le niveau à compléter. Ceci permettra de rendre encore mieux compte de la distribution de bordure positive et de diversifier les données réelles simulées.

Références

[AIS93] Rakesh Agrawal and Tomasz Imielinski and Arun N. Swami. Mining Association Rules between Sets of Items in Large Databases. *SIGMOD Conference*, pages 207-216, 1993.

[Ber07] P. Bergeret. Distribution de bases de transactions synthétiques : vers la prise en compte d'un distance entre les distributions des motifs fréquents. *Rapport de recherche 2007.* http ://liris.cnrs.fr/˜fflouvat/RapportPBergeret.pdf

[DDM05] D. Devaurs et F. De Marchi. Génération de bases de transactions synthétiques : vers la prise en compte des bordures. In *21èmes Journées "bases de données avancées" (BDA'05)*, pages 119-133, oct 2005.

[FIMI03] FIMI '03. *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, Melbourne, Florida, USA, November 19, 2003.

[FIMI04] FIMI '04. *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, Brighton, UK, November 1, 2004.

[FIMIRep] B. Goethals. Frequent Itemset Mining Implementations Repository, 2004. http ://fimi.cs.helsinki.fi

[FMP05] F.Flouvat, F.De Marchi, J.-M.Petit. A thorough experimental study of datasets for frequent itemsets. *5th IEEE International Conference on Data Mining (ICDM'05)*, Houston, USA, November 2005, pages 162-169.

[Fexp] F. Flouvat. Study of frequent itemsets datasets. http ://liris.cnrs.fr/˜fflouvat/resressus.html

[HMS01] David Hand, Heikki Mannila and Padhraic Smyth. *Principles of Data Mining. Adaptative Computation and Machine Learning.* MIT Press, Cambridge, Massachusetts, 2001

[MT97] H. Mannila and H. Toivonen. Levelwise Search and Borders of Theories in Knowledge Discovery. *Data Mining and Knowledge Discovery*, 1(3 : 241-258, 1997.

[OSDM05] Bart Goethals and Siegfried Nijssen and Mohammed Javeed Zaki. Open Source Data Mining : Workshop Report. *OSDM'05, Proceeding of the ACM SIGKDD Open Source Data Mining Workshop*, Chicago,USA 2005.

[RMZ03] G.Ramesh, W.A.Maniatty 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 Database Systems*, pages 284-295. ACM, June 2003.

[RMZ05] G.Ramesh, M.J.Zaki and W. Maniatty. Distribution-Based Synthetic Database Generation Techniques for Itemset Mining. *IDEAS*, 2005, pages 307-316.