Extraction de motifs séquentiels contextuels

Julien Rabatel, Sandra Bringay

Extraction de motifs séquentiels contextuels

Julien Rabatel[*,**], Sandra Bringay[*,***]

[*]LIRMM, Université Montpellier 2, CNRS
161 rue Ada, 34392 Montpellier Cedex 5, France
[**]Tecnalia, Cap Omega, Rond-point Benjamin Franklin - CS 39521
34960 Montpellier, France
[***]Dpt MIAP, Université Montpellier 3, Route de Mende
34199 Montpellier Cedex 5, France
{rabatel,bringay}@lirmm.fr

Résumé. Les motifs séquentiels traditionnels ne tiennent généralement pas compte des informations contextuelles fréquemment associées aux données séquentielles. Dans le cas des séquences d'achats de clients dans un magasin, l'extraction classique de motifs se focalise sur les achats des clients sans considérer leur catégorie socio-professionnelle, leur sexe, leur âge. Or, en considérant le fait qu'un motif séquentiel est spécifique à un contexte donné, un expert pourra adapter sa stratégie au type du client et prendre les décisions adéquates. Dans cet article, nous proposons d'extraire des motifs de la forme *« l'achat des produits A et B suivi de l'achat du produit C est spécifique aux jeunes clients »*. En mettant en valeur les propriétés formelles de tels contextes, nous développons un algorithme efficace d'extraction de motifs séquentiels contextuels. Les expérimentations effectuées sur un jeu de données réelles montrent les apports et l'efficacité de l'approche proposée.

1 Introduction

La découverte de motifs séquentiels présente un éventail important d'applications, telles que l'étude de comportement des utilisateurs, de données issues de capteurs, de puces ADN, etc. Les motifs séquentiels visent à extraire des ensembles d'items fréquemment associés au cours du temps. Par exemple, dans le but d'analyser les séquences d'achats de clients dans un magasin, un motif séquentiel peut être *« fréquemment, les clients achètent les produits A et B ensemble, puis achètent le produit C »*. Un tel motif apporte une information générale sur le comportement des clients. Cependant, les données traitées contiennent très souvent des informations additionnelles telles que l'âge ou le sexe des clients.

Les motifs séquentiels traditionnels ne permettent pas de tenir compte de ces informations additionnelles. Or, l'intérêt pour le décideur est immédiat car l'association entre un motif et un contexte lui permet d'adapter sa stratégie et ainsi de mieux coller à la réalité. Par exemple, l'expert pourra obtenir des réponses aux questions *« Quels sont les comportements spécifiques aux clients âgés ? »*, *« Existe-t-il des comportements spécifiques aux jeunes hommes ? »* ou encore *« Quels sont les comportements généraux, qui ne dépendent pas du contexte ? »*. Dans cet

article, nous proposons d'extraire des motifs de la forme *« acheter les produits A et B, puis le produit C est un comportement spécifique aux jeunes, alors que l'achat des produits B et D suivi par le produit E est spécifique aux personnes agées »*.

L'extraction de tels motifs est en revanche beaucoup plus difficile car il est indispensable de prendre en compte les différents niveaux de généralisation/spécialisation des contextes. Par exemple, le contexte correspondant aux *jeunes clients* est plus général que celui des *jeunes clients de sexe masculin*. Les différentes possibilités de contextes étant nombreuses, extraire les motifs séquentiels spécifiques à chacun d'eux est particulièrement coûteux.

La découverte de contextes plus ou moins généraux où un motif séquentiel est spécifique est un problème qui n'a, à notre connaissance, pas encore été étudié. La fouille de motifs séquentiels multidimensionnels telle que décrite dans Pinto et al. (2001) ou Plantevit et al. (2005) utilise également ce type d'informations additionnelles. Cependant, les motifs multidimensionnels extraits doivent être fréquents dans l'ensemble des données. Les motifs uniquement fréquents dans un contexte donné ne sont donc pas considérés. La même remarque s'applique pour Ziembiński (2007) qui définit une notion de contexte plus riche mais n'extrait que les motifs fréquents sur l'ensemble de la base. Un autre champ de recherche peut être relié à notre problématique : la découverte de motifs émergents. Introduite dans Dong et Li (1999), elle vise à extraire, parmi plusieurs classes de données, les motifs qui sont significativement plus fréquents dans une classe. Toutefois, peu d'approches considèrent les données séquentielles (Ji et al. (2007), Chan et al. (2003)) et ces travaux ne considèrent que les séquences spécifiques à une classe prédéfinie de données, sans prendre en compte les différents niveaux de généralisation/spécialisation possibles.

Dans cet article, nous proposons une description formelle des contextes et des motifs séquentiels contextuels. En mettant en lumière les propriétés de ces contextes, nous décrivons un algorithme visant à extraire de manière efficace ces motifs. Nous présentons l'évaluation menée sur des données réelles et montrons les performances de l'approche proposée.

La suite de l'article est organisée de la manière suivante. La section 2 présente les motifs séquentiels traditionnels et explique pourquoi ils ne sont pas adaptés pour manipuler les données contextuelles. Nous introduisons dans la section 3 comment les informations contextuelles sont considérées dans le but d'extraire des motifs pertinents. L'algorithme proposé est décrit dans la section 4. L'évaluation de notre approche est exposée dans la section 5. Enfin, nous concluons et discutons les perspectives de ces travaux dans la section 6.

2 Motifs séquentiels traditionnels

Nous décrivons dans cette section les motifs séquentiels traditionnels et soulignons le besoin de prendre en compte les informations contextuelles associées aux séquences.

Les motifs séquentiels traditionnels, introduits dans Agrawal et Srikant (1995), peuvent être considérés comme une extension du concept d'itemsets fréquents de Agrawal et al. (1993) en considérant les estampilles temporelles associées aux items. La fouille de motifs séquentiels vise à extraire des ensembles d'items fréquemment associés au cours du temps. En considérant l'étude des achats dans une boutique, un motif séquentiel pourrait par exemple être : «40 % des clients achètent une télévision, puis plus tard achètent un lecteur DVD».

La découverte de motifs séquentiels est formellement définie comme suit. Soit $\mathcal{X}$ un ensemble d'*items* distincts. Un *itemset* est un sous-ensemble d'items, noté $I = (i_1 i_2 ... i_n)$, i.e.,

pour $1 \leq j \leq n$, $i_j \in \mathcal{X}$. Une *séquence* est une liste ordonnée d'itemsets. Une séquence s est notée $\langle I_1 I_2 ... I_k \rangle$, où $I_i \subseteq \mathcal{X}$ pour $1 \leq i \leq n$.

Soit $s = \langle I_1 I_2 ... I_m \rangle$ et $s' = \langle I'_1 I'_2 ... I'_n \rangle$ deux séquences. La séquence s est une *sous-séquence* de s', noté $s \sqsubseteq s'$, si $\exists i_1, i_2, ... i_m$ avec $1 \leq i_1 < i_2 < ... < i_m \leq n$ tel que $I_1 \subseteq I'_{i_1}$, $I_2 \subseteq I'_{i_2}$, ..., $I_m \subseteq I'_{i_m}$.

Une *base de séquences* $\mathcal{B}$ est une relation $\mathcal{R}(ID, S)$, où un élément $id \in dom(ID)$ est un identifiant de séquence, et $dom(S)$ est l'ensemble des séquences. La *taille* de $\mathcal{B}$, notée $|\mathcal{B}|$, est le nombre de tuples dans $\mathcal{B}$. Un tuple $\prec id, s \succ$ *supporte* une séquence α si α est une sous-séquence de s, i.e., $\alpha \sqsubseteq s$. Le support d'une séquence α dans la base de séquences $\mathcal{B}$ est le nombre de tuples dans $\mathcal{B}$ supportant α, i.e., $sup_{\mathcal{B}}(\alpha) = |\{\prec id, s \succ | \ (\prec id, s \succ \in \mathcal{B}) \wedge (\alpha \sqsubseteq s)\}|$. Etant donné un nombre réel $minSupp$ le seuil de *support minimum*, tel que $0 < minSupp \leq 1$, une séquence α est un motif séquentiel dans la base de séquences $\mathcal{B}$ si la proportion de tuples dans $\mathcal{B}$ supportant α est supérieure à $minSupp$, i.e., $sup_{\mathcal{B}}(\alpha) \geq minSupp \cdot |\mathcal{B}|$. La séquence α est alors dite *fréquente dans* $\mathcal{B}$.

id	Age	Genre	Séquence
s_1	jeune	homme	$\langle (ad)(b) \rangle$
s_2	jeune	homme	$\langle (ab)(b) \rangle$
s_3	jeune	homme	$\langle (a)(a)(b) \rangle$
s_4	jeune	homme	$\langle (c)(a)(bc) \rangle$
s_5	jeune	homme	$\langle (d)(ab)(bcd) \rangle$
s_6	jeune	femme	$\langle (b)(a) \rangle$
s_7	jeune	femme	$\langle (a)(b)(a) \rangle$
s_8	jeune	femme	$\langle (d)(a)(bc) \rangle$
s_9	âgé	male	$\langle (ab)(a)(bd) \rangle$
s_{10}	âgé	male	$\langle (bcd) \rangle$
s_{11}	âgé	male	$\langle (bd)(a) \rangle$
s_{12}	âgé	femme	$\langle (e)(bcd)(a) \rangle$
s_{13}	âgé	femme	$\langle (bde) \rangle$
s_{14}	âgé	femme	$\langle (b)(a)(e) \rangle$

TAB. 1 – *Une base contextuelle de séquences.*

Exemple 1 : *Le tableau 1 présente une base de séquences $\mathcal{B}$ décrivant les achats effectués par des clients dans un magasin. Dans la première colonne, on trouve l'identifiant de chaque séquence. Ici, a, b, c, d, e sont des produits. Les colonnes Genre et Age sont des informations additionnelles relatives aux séquences. Elles ne sont pas considérées dans l'extraction de motifs séquentiels traditionnels. La taille de $\mathcal{B}$ est $|\mathcal{B}| = 14$. La première séquence décrit la séquence d'achats d'un client d'identifiant s_1 : il a acheté les produits a et d, puis le produit b.*

Dans la suite, nous fixons le support minimum $minSup$ à 0.5. Considérons la séquence $s = \langle (a)(b) \rangle$. Son support dans $\mathcal{B}$ est $sup_{\mathcal{B}}(s) = 8$. Ainsi, $sup_{\mathcal{B}}(s) \geq minSupp \cdot |\mathcal{B}|$, et s est un motif séquentiel dans $\mathcal{B}$.

Les informations contextuelles sont associées à une séquence de données. En considérant l'exemple précédent, les informations contextuelles disponibles sont l'âge et le genre des clients. Un contexte dans ce cas pourra être *jeune femme* ou encore *client âgé* (pour n'importe

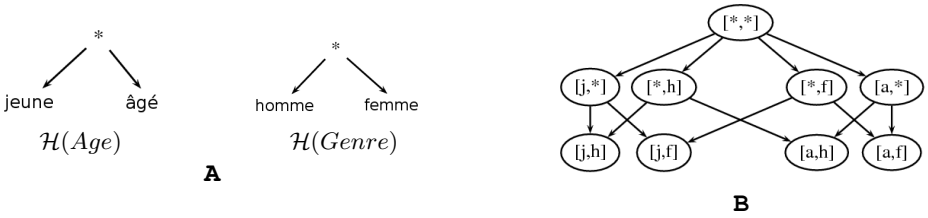


FIG. 1 – **A** - *Hiérarchies sur les dimensions Age et Genre.* **B** - *La hiérarchie de contextes* $\mathcal{H}$.

quel genre). Afin de comprendre les inconvénients posés par la fouille de motifs séquentiels traditionnels dans de telles données, considérons les deux exemples suivants.

Cas 1. La séquence $s = \langle (a)(b) \rangle$ est un motif séquentiel dans $\mathcal{B}$. Pourtant, ce motif semble spécifique aux jeunes clients : 7 *jeunes* clients sur 8 supportent cette séquence, contre seulement 1 client *âgé* sur 6.

Cas 2. La séquence $s' = \langle (bd) \rangle$ n'est pas un motif séquentiel (6 clients sur 14 la supportent) mais est fréquente pour les clients *âgés* : 5 sur 6 la supportent.

L'extraction de motifs séquentiels traditionnels peut ainsi mener à considérer certains comportements dépendants du contexte comme généraux (*cf.* Cas 1) alors qu'ils sont spécifiques à une sous-partie de la base. Au contraire, l'extraction traditionnelle peut également mener à ne pas les considérer comme fréquents parce que le contexte associé n'est lui-même pas fréquent (*cf.* Cas 2). Ainsi, dès lors que des informations contextuelles sont disponibles, leur prise en compte apporte une réelle valeur ajoutée pour les connaissances extraites.

3 Motifs séquentiels contextuels

Nous proposons dans cette section une description formelle de la notion de contexte, et définissons les notions nécessaires pour appréhender les motifs séquentiels contextuels.

3.1 Base contextuelle de séquences

Une *base contextuelle de séquences* est définie comme une relation $\mathcal{R}(ID, S, D_1, ...D_n)$, où $dom(S)$ est un ensemble de séquences et $dom(D_i)$ pour $1 \leq i \leq n$ est l'ensemble de toutes les valeurs possibles de D_i. D_1, D_2, ... D_n sont appelées les *dimensions contextuelles* de $\mathcal{CB}$. Un *tuple* $u \in \mathcal{CB}$ est noté $\prec id, s, d_1, ..., d_n \succ$.

Les valeurs de chaque dimension contextuelle peuvent être organisées sous la forme d'une hiérarchie. Ainsi pour $1 \leq i \leq n$, $dom(D_i)$ est étendu à $\mathcal{H}(D_i)$, où $\subseteq_{D_i}$ est un ordre partiel sur $\mathcal{H}(D_i)$ tel que $dom(D_i)$ est l'ensemble des éléments minimaux de $\mathcal{H}(D_i)$.

Exemple 2 : *Considérons les dimensions contextuelles Age et Genre. La figure 1-*A *présente un exemple de hiérarchie pour chacune d'elles. Dans cet exemple,* $\mathcal{H}(Age) = dom(Age) \cup \{*\}$, *où* $jeune \subseteq_{Age} *$ *et âgé* $\subseteq_{Age} *$.

Un *contexte* c dans $\mathcal{CB}$ est noté $[d_1, ...d_n]$, où $d_i \in \mathcal{H}(D_i)$. Si pour $1 \leq i \leq n$, $d_i \in dom(D_i)$, alors c est un *contexte minimal*.

Soit c_1 et c_2 deux contextes dans $\mathcal{CB}$, tels que $c_1 = [d_1^1, ..., d_n^1]$ et $c_2 = [d_1^2, ...d_n^2]$. Alors $c_1 \geq c_2$ si et seulement si $\forall i$ avec $1 \leq i \leq n$, $d_i^1 \supseteq_i d_i^2$. De plus, si $\exists i$ avec $1 \leq i \leq n$, tel que $d_i^1 \supset_i d_i^2$ alors $c_1 > c_2$. Dans ce cas, c_1 est *plus général* que c_2, et c_2 est *plus spécifique* que c_1.

Exemple 3 : *Dans la base contextuelle de séquences du tableau 1, il y a quatre contextes minimaux :* $[j, h]$, $[j, f]$, $[a, h]$, *et* $[a, f]$, *où j, a, h et f représentent respectivement jeune, âgé, homme et femme. Le contexte* $[*, *]$ *est plus général que* $[j, *]$. $[j, *]$ *et* $[*, h]$ *sont incomparables.*

L'ensemble de tous les contextes joint à l'ordre partiel $\geq$ constitue la hiérarchie de contextes, notée $\mathcal{H}$. Etant donné deux contextes c_1 et c_2 tels que $c_1 > c_2$, c_1 est un *ancêtre* de c_2, et c_2 est un *descendant* de c_1.

La figure 1-B montre une représentation visuelle de $\mathcal{H}$ pour les données fournies par le tableau 1, associées aux hiérarchies précédemment définies sur les dimensions *Age* et *Genre*.

Nous pouvons désormais considérer les tuples de $\mathcal{CB}$ conformément aux contextes définis plus tôt. Soit $u = \prec id, s, d_1, ...d_n \succ$ un tuple dans $\mathcal{CB}$. Le contexte c tel que $c = [d_1, ...d_n]$ est appelé le *contexte de* u. Notons que ce contexte est minimal, puisque $\forall i$ tel que $1 \leq i \leq n$, $d_i \in dom(D_i)$.

Soit u un tuple dans $\mathcal{CB}$ et c le contexte de u. Un contexte c' *contient* u (et u est contenu par c') si et seulement si $c' \geq c$.

Soit $c = [d_1, ...d_n]$ un contexte (minimal ou non) dans $\mathcal{CB}$, et $\mathcal{U}$ l'ensemble des tuples contenus par c. La *base de séquences de* c, notée $\mathcal{B}(c)$, est l'ensemble des tuples $\prec id, s \succ$ tels que $\exists u \in \mathcal{U}$ avec $u = \prec id, s, d_1, ...d_n \succ$. Nous définissons la *taille d'un contexte* c, notée $|c|$, comme la taille de sa base de séquences, i.e., $|c| = |\mathcal{B}(c)|$.

Exemple 4 : *Dans le tableau 1,* $\mathcal{B}([a, *]) = \{s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}\}$, *et* $|[a, *]| = 6$.

Soit un contexte c dans $\mathcal{CB}$. La décomposition de c dans $\mathcal{CB}$, notée $decomp(c)$, est l'ensemble non-vide $\{c_1, c_2, ..., c_n\}$ de contextes minimaux c tels que $c \geq c'$.
D'après la définition de $\mathcal{B}(c)$, la décomposition de c possède les propriétés suivantes :

$$1) \bigcap_{i=1}^{n} \mathcal{B}(c_i) = \emptyset\,; \quad 2) \bigcup_{i=1}^{n} \mathcal{B}(c_i) = \mathcal{B}(c)\,; \quad 3) |c| = |\mathcal{B}(c)| = \sum_{i=1}^{n} |c_i|.$$

Exemple 5 : *La décomposition de* $[j, *]$ *est* $\{[j, h], [j, f]\}$.

3.2 Motifs séquentiels contextuels

Nous avons montré dans la section précédente comment une base contextuelle de séquences peut être décomposée en s'appuyant sur les contextes. Désormais, considérons la définition d'un motif séquentiel dans de tels contextes.
Soit un contexte c et une séquence s.

Définition 1 : *s est un* ***motif séquentiel dans*** *c si et seulement si s est un motif séquentiel dans* $\mathcal{B}(c)$, *i.e., si* $sup_{\mathcal{B}(c)}(s) \geq minSup \cdot |c|$. *Par la suite, nous noterons* $sup_{\mathcal{B}(c)}(s)$ *par* $sup_c(s)$.

Nous souhaitons extraire les motifs pouvant être qualifiés de spécifiques à un contexte particulier. Nous définissons donc ici la notion de spécificité à un contexte.

Définition 2 : *s est* ***spécifique à c*** *(ou c-spécifique) si et seulement si :*

1. *s est un motif séquentiel dans c.*

2. *s est **général dans c**, i.e., s est un motif séquentiel dans tous les descendants de c dans la hiérarchie de contextes. Dans ce cas, s est dite **c-générale**.*
3. *il n'existe pas de contexte c', tel que $c' > c$ et s est c'-générale.*

Selon cette définition, un motif est c-spécifique s'il est fréquent dans tous les contextes descendants de c, et si c est le contexte le plus général qui respecte cette propriété.

Définition 3 : *Un **motif séquentiel contextuel** est un couple (c, s), tel que s est c-spécifique. (c, s) est alors **généré par** s.*

Exemple 6 : *Considérons la séquence $s = \langle(a)(b)\rangle$. Les supports de s dans les contextes minimaux de $\mathcal{CB}$ sont présentés ci-dessous.*

	$[j,h]$	$[j,f]$	$[a,h]$	$[a,f]$
$\langle(a)(b)\rangle$	***5/5***	***2/3***	*1/3*	*0/3*

*s est fréquent dans $[j,*]$ (7 jeunes sur 8 la supportent), ainsi que dans ses descendants $[j,h]$ et $[j,f]$. De plus, s n'est pas $[*,*]$-générale (car il existe des descendants de $[*,*]$ dans lesquels s n'est pas fréquent). Par conséquent, s est $[j,*]$-spécifique et $([j,*],s)$ est un **motif séquentiel contextuel**.*

3.3 Extraction des motifs séquentiels contextuels

Les concepts liés aux motifs séquentiels contextuels étant désormais définis, nous nous intéressons à leur extraction. Une approche naïve consiste à extraire les motifs séquentiels indépendamment dans chaque élément de la hiérarchie des contextes, puis pour chaque contexte à éliminer les motifs non-spécifiques. Cette approche soulève deux difficultés :

- **Les contextes à fouiller sont nombreux.** En effet, le nombre d'éléments d'une hiérarchie de contextes est $\prod_{i=1}^{n} |\mathcal{H}(D_i)|$, où $D_1, ...D_n$ sont les dimensions contextuelles. En comparaison, le nombre de contextes minimaux est $\prod_{i=1}^{n} |dom(D_i)|$.
- **Eliminer les motifs séquentiels n'ayant pas les propriétés requises est coûteux**. En effet, vérifier qu'un motif est spécifique à un contexte c donné nécessite de contrôler sa fréquence dans tous les autres contextes de la hiérarchie.

Afin de surmonter ces difficultés, nous étudions les propriétés de la hiérarchie de contextes et montrons que les motifs séquentiels contextuels peuvent être générés en considérant uniquement les motifs séquentiels des contextes minimaux. Les propriétés de la décomposition d'un contexte impliquent le lemme suivant.

Lemme 1 : *Soit un contexte c, tel que $decomp(c) = \{c_1, c_2, ..., c_n\}$. Si $\forall i \in 1, ..., n$, s est un motif séquentiel dans c_i (respectivement n'est pas un motif séquentiel), alors s est un motif séquentiel dans c (respectivement n'est pas un motif séquentiel dans c). De plus, s est un motif séquentiel (respectivement n'est pas un motif séquentiel) dans les descendants de c.*

Démonstration : *Pour tout c_i tel que $i \in \{1, ..n\}$, $sup_{c_i}(s) \geq minSup \cdot |c_i|$. Cela signifie que $\sum_{i=1}^{k} sup_{c_i}(s) \geq \sum_{i=1}^{n} minSup \cdot |c_i|$. Cependant, $\sum_{i=1}^{n} minSup \cdot |c_i| = minSup \cdot \sum_{i=1}^{n} |c_i| = minSup \cdot |c|$. Comme $\sum_{i=1}^{k} sup_{c_i}(s) = sup_c(s)$, $sup_c(s) \geq minSup \cdot |c|$.*

Soit un contexte c' tel que $c > c'$. Alors $decomp(c') \subseteq decomp(c)$, i.e., s est un motif séquentiel dans chaque élément de $decomp(c')$. Par application du résultat précédent, s est un motif séquentiel dans c'.

Un raisonnement similaire est appliqué si s n'est un motif séquentiel dans aucun des éléments de $decomp(c)$. □

Le lemme 1 est un résultat important car il nous permet de redéfinir la notion de c-spécificité en ne tenant compte que de la décomposition des contextes de la hiérarchie.

Dans la suite de cette section, nous notons $\mathcal{F}$ l'ensemble des contextes minimaux dans lesquels s est fréquent. En exploitant le lemme 1, nous constatons que s est c-spécifique si et seulement si *(i)* $decomp(c) \subseteq \mathcal{F}$ et *(ii)* il n'existe pas de contexte c' tel que $c' > c$ et $decomp(c') \subseteq \mathcal{F}$. L'ensemble des contextes vérifiant ces conditions est appelé la *couverture* de $\mathcal{F}$ et noté $cov(\mathcal{F})$. Nous montrons dans la section 4 comment calculer la couverture de $\mathcal{F}$ à partir de la hiérarchie de contextes.

Exemple 7 : *Soit $\mathcal{F} = \{[j,h],[j,f],[a,f]\}$, alors $cov(\mathcal{F}) = \{[j,*],[*,f]\}$ et $([j,*],s)$ et $([*,f],s)$ sont les motifs séquentiels contextuels générés par s.*

Théorème 1 : *Soit S l'ensemble des séquences fréquentes dans au moins un contexte minimal. L'ensemble des motifs séquentiels contextuels est l'ensemble de tous les couples (c,s) où $s \in S$ et (c,s) est généré par s.*

Démonstration : *Ce résultat est une conséquence immédiate de la définition d'une séquence c-spécifique. En effet, si s n'est fréquent dans aucun contexte minimal, i.e., $\mathcal{F} = \emptyset$, alors il n'est fréquent dans aucun élément de la hiérarchie de contextes (voir lemme 1) et il n'existe aucun contexte c tel que s est c-spécifique. Ainsi, tout motif séquentiel contextuel est généré par une séquence qui est fréquente dans au moins un contexte minimal.* □

Le théorème 1 est essentiel dans le problème de l'extraction de motifs séquentiels contextuels. En effet, il assure que tous les motifs séquentiels contextuels peuvent être déduits des motifs séquentiels des contextes minimaux. Dans la section 4, nous nous appuyons sur les propriétés des motifs séquentiels contextuels pour proposer un algorithme d'extraction efficace.

4 Algorithme

L'algorithme d'extraction de motifs séquentiels contextuels proposé est basé sur l'approche PrefixSpan dédiée à l'extraction de motifs séquentiels traditionnels (Pei et al. (2004)). L'exemple suivant décrit le principe de PrefixSpan en l'appliquant sur la base de séquences du tableau 1, avec un support minimum fixé à *0.5*.

Exemple 8 : *Un premier parcours de la base de séquences extrait tous les motifs séquentiels de la forme $\langle(i)\rangle$, où i est un item. Dans l'exemple, on obtient les motifs : $\langle(a)\rangle$, $\langle(b)\rangle$, $\langle(d)\rangle$. On ne trouve pas les motifs $\langle(c)\rangle$ et $\langle(e)\rangle$ qui ne sont pas fréquents.*

*Par conséquent, l'ensemble des motifs séquentiels dans $\mathcal{B}$ peut être partitionné en sous-ensembles, chacun d'eux étant l'ensemble des motifs séquentiels ayant $\langle(i)\rangle$ pour préfixe. PrefixSpan repose sur le fait que ces sous-ensembles peuvent être extraits des **bases projetées** de chaque préfixe, i.e., pour chaque $\langle(i)\rangle$. Une base projetée contient, pour chaque séquence de $\mathcal{B}$, sa sous-séquence contenant tous les items fréquents suivant la première occurrence du pré-*

fixe donné. Une telle sous-séquence est appelée ***postfixe****. Si le premier item x du postfixe est présent dans le même itemset que le dernier item du préfixe, le postfixe est noté $\langle(\_x...)...\rangle$.*

Considérons le motif séquentiel $\langle(a)\rangle$. La base projetée de $\langle(a)\rangle$ contient 11 postfixes : $\langle(\_d)(b)\rangle$, $\langle(\_b)(b)\rangle$, $\langle(a)(b)\rangle$, $\langle(bc)\rangle$ etc. On extrait ensuite tous les items i tels que $\langle(ai)\rangle$ ou $\langle(a)(i)\rangle$ soit fréquent. Ici, b est un de ces items, car $\langle(a)(b)\rangle$ est fréquent. Le processus peut ainsi continuer en retournant $\langle(a)(b)\rangle$, puis en l'utilisant comme un nouveau préfixe.

Dans la suite de cette section, nous présentons l'algorithme d'extraction de motifs séquentiels contextuels. Notre approche peut être décomposée en deux étapes principales :

1. À partir d'une base contextuelle de séquences, nous extrayons toutes les séquences qui sont fréquentes dans au moins un contexte minimal. Tous les motifs séquentiels contextuels peuvent en effet être générés à partir de cet ensemble ;
2. D'après l'ensemble de séquences obtenues à l'étape 1, nous générons l'ensemble des motifs séquentiels contextuels.

Les étapes décrites ci-dessus sont présentées dans l'algorithme 1, qui tire parti de l'algorithme 2. En prenant en entrée une base contextuelle de séquences $\mathcal{CB}$, un seuil de support minimum $minSup$, et une hiérarchie de contextes $\mathcal{H}$, l'algorithme retourne les motifs séquentiels contextuels de $\mathcal{CB}$.

Extraction des motifs séquentiels dans les contextes minimaux. La première étape de l'algorithme est l'extraction des motifs séquentiels dans les contextes minimaux, chacun étant associé à l'ensemble des contextes minimaux dans lequel il est fréquent. Cette étape est effectuée en utilisant le principe de PrefixSpan : en prenant pour préfixe une séquence s, l'algorithme construit (méthode $ConstruitBaseProjetée$) et parcourt la base projetée correspondante (méthode $ParcourtBase$) afin de trouver les items i qui peuvent être assemblés pour former un nouveau motif séquentiel s'. Puis, le processus continue avec le nouveau préfixe s'. Les méthodes étant proches de PrefixSpan, nous ne détaillons pas ici $ParcourtBase$ et $ConstruitBaseProjetée$, mais en présentons seulement les principales caractéristiques.

- $ParcourtBase(\mathcal{CB})$: la différence principale repose sur le fait que le support de i est calculé pour chaque contexte minimal de la base projetée. Ainsi, cette méthode retourne l'ensemble des couples $(i, \mathcal{F}_i)$, où i est un item et $\mathcal{F}_i$ est l'ensemble non vide de contextes minimaux où i est fréquent.
- $ConstruitBaseProjetée(s, \mathcal{F}_i, \mathcal{CB})$: ici, seules les séquences contenues dans les contextes de $\mathcal{F}_i$ sont considérées pour la construction de la base projetée de s. En effet, si s n'est pas fréquent dans un contexte c, alors les séquences construites à partir de s ne sont également pas fréquentes dans c.

Génération des motifs séquentiels contextuels. Un motif séquentiel s est extrait avec l'ensemble de contextes minimaux où il est fréquent. En nous appuyant sur cet ensemble, nous déduisons les motifs séquentiels générés par s, i.e., l'ensemble de (c, s) où c est un contexte tel que s est c-spécifique. Ceci est réalisé par $Couverture(\mathcal{F}, \mathcal{H})$, décrit dans l'algorithme 2. Cet algorithme repose sur un parcourt ascendant de la hiérarchie de contextes (i.e., des feuilles vers la racine), dans le but de collecter les contextes les plus généraux dont la décomposition est un sous-ensemble de $\mathcal{F}$, i.e., où un motif séquentiel est spécifique (*cf.* section 3).

Algorithm 1 FouilleContextuelle

ENTRÉES: une base contextuelle de séquences $\mathcal{CB}$, un support minimum $minSup$, une hiérarchie de contextes $\mathcal{H}$.
Appelle $auxiliaireFouilleContextuelle(\langle\rangle, \mathcal{CB}, \mathcal{H})$

Routine $auxiliaireFouilleContextuelle(s, \mathcal{CB}, \mathcal{H})$
ENTRÉES: une séquence s ; la s-base projetée $\mathcal{CB}$, une hiérarchie de contextes $\mathcal{H}$.
$I = ParcourtBase(\mathcal{CB})$;
pour tout couple $(i, \mathcal{F}_i) \in I$ **faire**
s' est la séquence telle que i est assemblé avec s ;

```
/* Génération des motifs séquentiels contextuels */
```
$C = Couverture(\mathcal{F}_i, \mathcal{H})$
pour tout $c \in C$ **faire**
affiche (s', c) ;
fin pour

$\mathcal{CB}' = ConstruitBaseProjetée(s, \mathcal{CB})$;
appelle $auxiliaireFouilleContextuelle(s', \mathcal{CB}', \mathcal{H})$
fin pour

5 Expérimentations

Description des données. Les expérimentations ont été menées sur environ 100000 commentaires d'utilisateurs sur des produits du site *amazon.com*, avec l'objectif d'étudier le vocabulaire utilisé en fonction du type de commentaire. Ce jeu de données est une partie de celui utilisé dans Jindal et Liu (2008). Les commentaires, en anglais, ont été lemmatisés [1] et grammaticalement filtrés afin de supprimer les termes jugés inintéressants, en utilisant l'outil *tree tagger* de Schmid (1994). Nous avons conservé les verbes (mis à part les verbes modaux et le verbe «être»), les noms, les adjectifs et les adverbes. La base de séquences a été construite suivant les principes suivants : chaque commentaire est une séquence, chaque phrase est un itemset (i.e., l'ordre des mots dans une phrase n'est pas considéré), et chaque mot lemmatisé est un item. Un motif séquentiel est $\langle(eat\ mushroom)(hospital)\rangle$, signifiant que fréquemment, une phrase contient les mots *eat* et *mushroom* et une des phrases suivantes contient *hospital*.

Chaque commentaire est associé aux dimensions contextuelles suivantes :

- le type de *produit* (*Book*, *DVD*, *Music* ou *Video*).
- la *note* (à l'origine, une valeur numérique r entre 0 et 5). Pour ces expérimentations, r a été traduit en valeur qualitatives : *Bad* (si $0 \leq r < 2$), *Neutral* (si $2 \leq r \leq 3$), et *Good* (si $3 < r \leq 5$).
- la *réaction* : pourcentage de réactions positives sur ce commentaire [2], i.e., *0-25%*, *25-50%*, *50-75%* or *75-100%*.

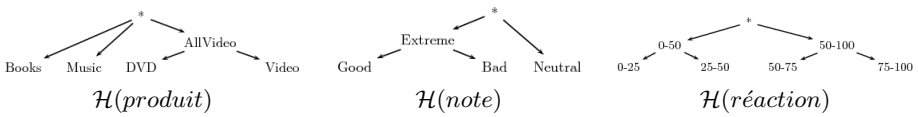
Nous définissons les hiérarchies sur les dimensions contextuelles comme décrites dans la figure 2. Le nombre de contextes est $|\mathcal{H}(produit)| \times |\mathcal{H}(note)| \times |\mathcal{H}(réaction)| = 6 \times 5 \times 7 = 210$, le nombre de contextes minimaux est $dom(produit) \times dom(note) \times dom(réaction) = 4 \times 3 \times 4 = 48$.

1. i.e., les différentes formes d'un mot ont été regroupées sous la forme d'un item unique. Par exemple, les différentes formes du verbe *être* (est, sont, était, été, etc.) sont toutes retournées en «être».
2. Sur *amazon.com*, chaque utilisateur peut poster sa réaction sur un commentaire.

Algorithm 2 Couverture($\mathcal{F}$, $\mathcal{H}$)

ENTRÉES: Un ensemble de contextes minimaux $\mathcal{F}$, une hiérarchie de contextes $\mathcal{H}$.
Soit $C = \emptyset$
Soit $\mathcal{L}$ l'ensemble des feuilles de $\mathcal{H}$
pour tout $l \in \mathcal{L}$ **faire**
$C = C \cup auxiliaireCouverture(l, \mathcal{F}, \mathcal{H})$
fin pour
retourne C la couverture de $\mathcal{F}$ dans $\mathcal{H}$

Routine $auxiliaireCouverture(c, \mathcal{F}, \mathcal{H})$
ENTRÉES: Un contexte c, un ensemble de contextes minimaux $\mathcal{F}$, une hiérarchie de contextes $\mathcal{H}$.
Soit $C = \emptyset$
si $decomp(c) \subseteq \mathcal{F}$ **alors**
pour tout p parent de c dans $\mathcal{H}$ **faire**
$C = C \cup auxiliaireCouverture(p, \mathcal{F}, \mathcal{H})$
fin pour
si $C = \emptyset$ **alors**
$C = \{c\}$
finsi
finsi
retourne C



$\mathcal{H}(produit)$ $\mathcal{H}(note)$ $\mathcal{H}(réaction)$

FIG. 2 – *Hiérarchies sur les dimensions contextuelles.*

Notons que le domaine des dimensions contextuelles est enrichi avec de nouvelles valeurs. Par exemple, la hiérarchie $\mathcal{H}(note)$ contient une valeur $Extreme$, qui permettra par la suite d'obtenir les motifs spécifiques aux opinions extrêmes, qu'elles soient positives ou négatives.

Résultats. Toutes les expérimentations ont été effectuées sur un système équipé de 16GB de mémoire centrale et d'un processeur cadencé à 3GHz. Les méthodes sont implémentées en étendant une implémentation Java de PrefixSPan décrite dans Fournier-Viger et al. (2008).

La figure 3-A montre le passage à l'échelle de notre approche. Le temps d'exécution augmente presque linéairement avec la taille de la base de séquences, que nous faisons évoluer de 12400 à 99834 tuples (i.e., le jeu de données complet) [3].

La figure 3-C montre le temps d'exécution du processus global (i.e., l'extraction des motifs séquentiels dans les 48 contextes minimaux et la génération des motifs séquentiels contextuels) tandis que la figure 3-D présente le temps d'exécution en millisecondes pour la génération des motifs séquentiels contextuels uniquement. Cette génération est extrêmement rapide. En comparaison, l'approche naïve décrite dans la section 3 nécessiterait d'extraire les motifs séquentiels dans les 210 contextes (au lieu des 48 minimaux). Revenons sur les valeurs ajoutées aux hiérarchies sur les dimensions contextuelles (par exemple, la valeur $Extreme$ de $\mathcal{H}(note)$). L'enrichissement de la hiérarchie de contextes par l'ajout de telles valeurs ne modifie en rien

3. Pour faire varier le nombre de tuples, nous avons aléatoirement sélectionné des tuples dans chaque contexte minimal, dans le but de garder la même répartition de tuples que sur la base entière.

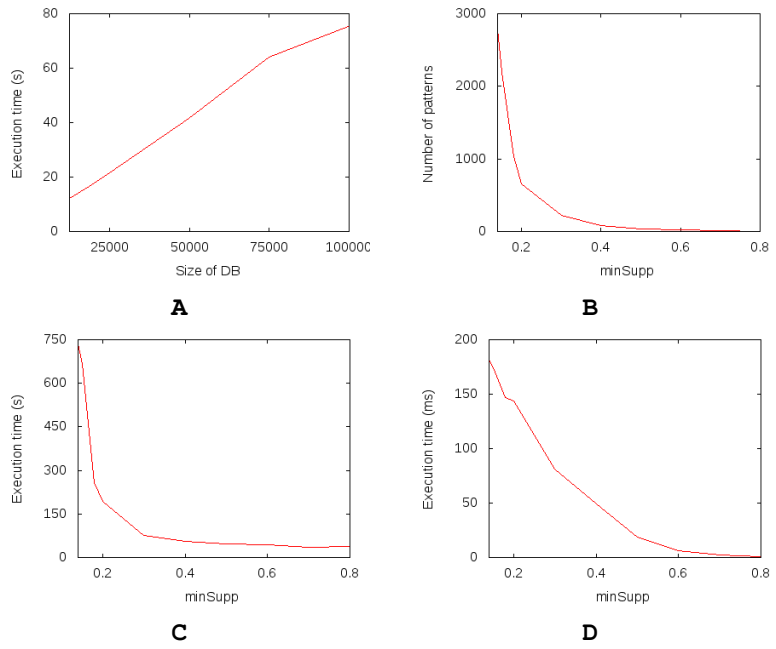


FIG. 3 – **A** - *Temps d'exécution en fonction de la taille de $\mathcal{CB}$ (avec $minSupp = 0.3$).* **B**- *Nombre de motifs séquentiels contextuels en fonction de $minSupp$.* **C** - *Temps d'exécution global en fonction de $minSupp$ (en secondes).* **D** - *Temps d'exécution pour la génération des motifs séquentiels contextuels en fonction de $minSupp$ (en millisecondes).*

le nombre de contextes minimaux, et n'a par conséquent aucune influence sur le temps d'extraction des motifs séquentiels dans ceux-ci. Or, le temps de génération des motifs séquentiels contextuels étant minime, l'impact sur le temps global du processus est négligeable.

Considérons maintenant les résultats obtenus. Parmi les 2193 motifs séquentiels extraits pour $minSup = 0.15$ (*cf.* figure 3-B), 13 sont $[*, *, *]$-spécifiques. Seulement 0.6% des motifs extraits sont indépendants du contexte, i.e., ils sont fréquents peu importe le contexte considéré. Ce fait souligne le besoin de prise en compte du contexte dans le processus de fouille.

6 Conclusions et perspectives

Dans cet article, nous avons soulevé le problème de la fouille de motifs séquentiels contextuels. Nous avons formellement défini les concepts nécessaires et les propriétés essentielles utilisées pour proposer un algorithme efficace d'extraction de tels motifs. Ces travaux ouvrent de nombreuses perspectives. Dans un premier temps, les travaux futurs incluront des expérimentations sur différents jeux de données réels, ainsi qu'une comparaison avec une approche naïve ne bénéficiant pas des propriétés formelles mises à jour. Nous étudierons également comment les résultats obtenus peuvent être exploités pour la classification. Par exemple, pour classer un client C en fonction de sa séquence d'achats, nous pourrons exploiter les motifs

contextuels, et ainsi fournir une connaissance du type : C a été classé comme *jeune* avec une confiance de 95%, mais comme *jeune femme* avec une confiance de 60% seulement.

Références

Agrawal, R., T. Imieliński, et A. Swami (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec. 22*(2).

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In P. S. Yu et A. S. P. Chen (Eds.), *Eleventh International Conference on Data Engineering*. IEEE Computer Society Press.

Chan, S., B. Kao, C. Yip, et M. Tang (2003). Mining emerging substrings. In *Database Systems for Advanced Applications, 2003.(DASFAA 2003). Proceedings. Eighth International Conference on*.

Dong, G. et J. Li (1999). Efficient mining of emerging patterns : discovering trends and differences. In *KDD '99 : Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA. ACM.

Fournier-Viger, P., R. Nkambou, et E. Nguifo (2008). A knowledge discovery framework for learning task models from user interactions in intelligent tutoring systems. *MICAI 2008 : Advances in Artificial Intelligence*.

Ji, X., J. Bailey, et G. Dong (2007). Mining minimal distinguishing subsequence patterns with gap constraints. *Knowledge and Information Systems 11*(3).

Jindal, N. et B. Liu (2008). Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*. ACM.

Pei, J., J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, et M. Hsu (2004). Mining sequential patterns by pattern-growth : the PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering 16*(11).

Pinto, H., J. Han, J. Pei, K. Wang, Q. Chen, et U. Dayal (2001). Multi-dimensional sequential pattern mining. In *Proceedings of the tenth international conference on Information and knowledge management*. ACM.

Plantevit, M., Y. W. Choong, A. Laurent, D. Laurent, et M. Teisseire (2005). M^2SP : Mining sequential patterns among several dimensions. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho, et J. Gama (Eds.), *PKDD*, Volume 3721 of *Lecture Notes in Computer Science*. Springer.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Volume 12. Citeseer.

Ziembiński, R. (2007). Algorithms for context based sequential pattern mining. *Fundamenta Informaticae 76*(4), 495–510.

Summary

Traditional sequential patterns do not consider contextual information oftenly associated with sequential data. In this paper, we propose to mine patterns of the form *"the purchase of products A and B, followed by the purchase of product C is specific to young customers"*. We present algorithm to extract contextual sequential patterns and conduct experiments on a real-world dataset to show the interesting intake and the efficiency of the proposed approach.