Fouille de données multidimensionnelles : différentes stratégies pour prendre en compte la mesure

Marc Plantevit, Anne Laurent, Maguelonne Teisseire

Fouille de Données Multidimensionnelles : Différentes Stratégies pour Prendre en Compte la Mesure

Marc Plantevit, Anne Laurent, Maguelonne Teisseire

LIRMM, Université Montpellier 2, CNRS,
161 Rue Ada 34392 Montpellier, France
{plantevi, laurent, teisseire}@lirmm.fr

Résumé. Les entrepôts de données contiennent de grosses masses de données historisées stockées à des fins d'analyse. Des techniques d'extraction de motifs séquentiels multidimensionnels ont été développées afin de mettre en exergue des corrélations entre des positions sur des dimensions au cours du temps. Même si ces méthodes offrent une meilleure appréhension des données sources en prenant en compte certaines spécificités des cubes de données (e.g. multidimensionnalité, hiérarchies, relation d'ordre), aucune méthode ne permet de prendre directement en compte la valeurs des agrégats (mesure) dans l'extraction des motifs. Dans cet article, nous définissons deux méthodes de comptage du support d'une séquence multidimensionnelle en s'appuyant sur les valeurs des agrégats des cellules qui supportent cette séquence. Des expérimentations sont décrites et montrent l'intérêt de notre proposition.

1 Introduction

Il est largement reconnu que la technologie OLAP fournit de puissants outils d'analyse dans le but d'extraire des connaissances utiles dans d'importants volumes de données hétérogènes et distribués. Plusieurs avantages viennent confirmer le potentiel d'un tel modèle d'analyse. En effet, OLAP a l'avantage de pouvoir représenter de manière assez naturelle les ensembles de données de la vie réelle, qui sont multidimensionnels, définis sur plusieurs niveaux de hiérarchies, et fortement corrélés. Ainsi, les données peuvent être analysées suivant plusieurs niveaux d'agrégation à l'aide des nombreux opérateurs OLAP (roll-up, drill-down, etc.) ou par des requêtes complexes (top-k, iceberg, gradient). Enfin, l'une des principales forces d'OLAP est son intégration avec des outils plus complexes d'analyse issus des statistiques, de l'analyse des séries temporelles et de la fouille de données. A partir de ce dernier, beaucoup de travaux visent à coupler OLAP à des outils et des algorithmes de fouilles de données à partir de la proposition de Han (1998). Cette proposition consiste à combiner la puissance d'OLAP à l'efficacité des algorithmes de fouilles de données permettant de découvrir des connaissances intéressantes dans de vastes volumes de données (e.g., l'ensemble des cellules d'un cube de données).

Dans ce contexte, l'extraction de motifs séquentiels permet de mettre en exergue des corrélations entre événements suivant leur chronologie d'apparition Agrawal et Srikant (1995). Extraire de tels motifs est particulièrement adapté dès qu'on souhaite découvrir des tendances

suivant une relation d'ordre (e.g., le temps). Ainsi, ils montrent leur efficacité dans de nombreux contextes (*e.g.* comportement des utilisateurs Srivastava et al. (2000), web log mining Pei et al. (2000), découverte de motifs dans des séquences de protéines Yang et al. (2002), sécurité, et musique Hsu et al. (2001)). Toutefois, cette problématique est très difficile du fait que l'espace de recherche associé à ce problème est très important. Pour surmonter cette difficulté, des algorithmes ont été développés, basés sur des heuristiques comme la génération par niveaux (Apriori) Agrawal et Srikant (1995); Masseglia et al. (1998); Zaki (2001); Ayres et al. (2002) ou en effectuant une recherche gloutonne à l'aide de multiples projections de la base de données Pei et al. (2004).

Dernièrement ces approches ont été étendues afin d'extraire des motifs séquentiels multidimensionnels Pinto et al. (2001), Plantevit et al. (2005), Yu et Chen (2005). Ces approches cherchent à découvrir des motifs plus intéressants en prenant en compte à la fois le temps et plusieurs dimensions d'analyse. Plantevit et al. (2006) propose de prendre en compte les hiérarchies dans l'extraction des motifs séquentiels multidimensionnels.

Même si les approches décrites précédemment s'attaquent à certaines spécifités inhérentes à OLAP comme la multidimensionnalité et la présence de hiérarchies, il n'y a pas de proposition qui tente de prendre directement en compte la mesure. Il existe de nombreux travaux permettant de discrétiser cette dimension numérique. Toutefois, ces travaux nécessitent un prétraitement des données, les algorithmes d'extraction de motifs séquentiels multidimensionnels étant lancés sur les données discrétisées. Aucune approche ne propose de prendre directement en compte le caractère numérique de la mesure sans passer par une étape de prétraitement.

Dans cet article, nous proposons de prendre en compte la mesure pour calculer le support des séquences multidimensionnelles. Le support d'une séquence permet de déterminer si celle-ci est fréquente ou non. La valeur d'agrégat d'une cellule peut être vue comme un « précalcul » du support d'une séquence. Il serait donc judicieux d'utiliser pleinement ces valeurs d'agrégats dans l'extraction de motifs séquentiels multidimensionnels. Nous proposons ainsi deux nouvelles définitions du support d'une séquence multidimensionnelle.

Le reste de l'article est organisé de la façon suivante. Dans la section 2, nous présentons les travaux actuels sur l'extraction de motifs séquentiels multidimensionnels. Nous montrons également les limites de ces approches lorsqu'elles doivent considérer les mesures des cellules dans un cube de données. Dans la section 3, nous présentons les travaux sur l'extraction de motifs qui essaient de prendre en compte des attributs numériques. Dans la section 4, nous proposons deux nouvelles définitions du support d'une séquence multidimensionnelle qui s'appuient sur la valeur des agrégats. Nous rapportons des expérimentations menées sur données synthétiques dans la section 5.

2 Extraction de motifs séquentiels multidimensionnels

Dans cette section, nous présentons les approches permettant l'extraction des motifs séquentiels multidimensionnels. Nous illustrons aussi les limites de ses approches dues à la non-prise en compte du caractère numérique de la mesure.

Combiner plusieurs dimensions d'analyse permet d'extraire des connaissances qui décrivent mieux les données. Dans Pinto et al. (2001) les auteurs sont les premiers à rechercher des motifs séquentiels multidimensionnels. Ainsi, les achats ne sont plus décrits en fonction des seuls date et identifiant du client, mais en fonction d'un ensemble de dimensions telles que

Type de consommateur, Ville, Age. Cette approche permet d'extraire des séquences d'items sur la dimension *produits* et de les caractériser à l'aide des informations fréquentes sur les clients (« *Patterns* ») qui tendent à supporter les séquences. Cette méthode ne permet pas d'avoir des séquences où plusieurs patterns sont présents. Elle ne permet donc pas d'extraire des connaissances de la forme : $\langle\{(business, *, *, a)(*, chicago, *, b)\}, \{(*, *, young, c)\}\rangle$ alliant différents *patterns* multidimensionnels.

L'approche proposée par Plantevit et al. (2005) permet, quant à elle, l'extraction de motifs séquentiels multidimensionnel *inter pattern*. Nous décrivons plus en détail les concepts associés dans la suite.

Dans Yu et Chen (2005), les auteurs proposent d'étendre la recherche de motifs séquentiels au contexte des bases de données décrivant les informations au moyen de plusieurs attributs. Cependant cette approche est restreinte au cas particulier où les dimensions étudiées entretiennent entre elles un très fort lien. En effet, ces dimensions sont organisées en hiérarchie. Ainsi, dans l'exemple pris par les auteurs, les différentes dimensions sont liées au comportement d'internautes dont les visites de pages sont organisées en transactions (dimension 1), elles-mêmes organisées en sessions (dimension 2), elles-mêmes organisées en jours (dimension 3). Ces différentes dimensions sont imbriquées au sein des motifs trouvés et il est impossible de retrouver les valeurs fréquentes le long de ces dimensions, celles-ci n'intervenant que pour organiser le temps de manière hiérarchique.

Nous pouvons encore citer les travaux de de Amo et al. (2004) qui proposent une approche basée sur la logique temporelle du premier ordre pour l'extraction de motifs séquentiels multidimensionnels, Lee (2005) propose également une nouvelle méthode de génération des séquences multidimensionnelles présentes dans des bases de transactions.

Nous détaillons ici les concepts relatifs à Plantevit et al. (2005). Considérons une base de données DB définie sur un ensemble de dimensions $\mathcal{D}$. Afin de permettre à l'utilisateur une plus grande liberté dans le choix des différents paramètres de l'extraction, les auteurs proposent une partition de $\mathcal{D}$ en quatre sous-ensembles :

- D_t pour les dimensions "*temporelles*", l'ensemble des dimensions permettant d'introduire une relation d'ordre en les événements (*e.g.* $temps$) ;
- D_A pour les dimensions d'*analyse*, l'ensemble des dimensions sur lesquelles les corrélations sont extraites (les dimensions décrivant les motifs extraits) ;
- D_R pour les dimensions de *référence*, l'ensemble des dimensions permettant de calculer le support d'une séquence et donc de déterminer si elle est fréquente ou non ;
- D_I pour les dimensions *ignorées*, l'ensemble des dimensions qui ne sont pas prises en compte durant l'extraction des motifs séquentiels multidimensionnels.

Chaque nuplet $c = (d_1, \ldots, d_n)$ peut ainsi s'écrire $c = (i, r, a, t)$ où i est la restriction sur D_I de c, r sa restriction sur D_R, a sa restriction sur D_A, et t sa restriction sur D_t.

Etant donnée une base de données DB, on appelle *bloc* l'ensemble des n-uplets ayant la même valeur r sur D_R. L'ensemble des blocs de DB est noté B_{DB,D_R}. Ainsi chaque bloc B_r de B_{DB,D_R} est décrit pas le tuple tuple r qui le définit.

Durant l'extraction de motifs séquentiels multidimensionnels, l'ensemble D_R identifie les blocs de la base de données qui doivent être considérée pour calculer le support d'une séquence. C'est pour cette raison que cet ensemble est nommé *référence*. Remarquons que pour les motifs séquentiels "classiques" et l'approche de Pinto et al. (2001), cet ensemble est un singleton (cid dans Pinto et al. (2001)) alors que dans Plantevit et al. (2005) la cardinalité de cet ensemble est

supérieure ou égale à 1. D'autre part, l'ensemble D_A décrit les dimensions d'*analyse*, ainsi les motifs définis sur ces dimensions seront découverts par un algorithme d'extraction de motifs séquentiels multidimensionnels. Pour les motifs séquentiels classiques, une seule dimension d'analyse est considérée, correspondant par exemple aux produits vendus ou aux pages web visitées. Enfin, l'ensemble D_I décrit les dimensions *ignorées* qui ne sont ni requises pour définir la relation d'ordre ou les motifs extraits, ni pour identifier les blocs.

CID	Date	City	Customer Informations		Product
			$Cust\text{-}Grp$	$Cust\text{-}Age$	
C_1	1	NY	$Educ.$	$Middle$	A
C_1	1	NY	$Educ.$	$Middle$	B
C_1	2	LA	$Educ$	$Middle$	C
C_2	1	SF	$Prof.$	$Middle$	A
C_2	2	SF	$Prof.$	$Middle$	C
C_3	1	DC	$Business$	$Retired$	A
C_3	1	LA	$Business$	$Retired$	B

FIG. 1 – DB *: Base de Données « Exemple »*

Afin d'illustrer les différents définitions, nous considérons une société de vente en ligne stockant les opérations de ses clients dans une base de données. La figure Fig. 1 représente un morceau de cette base de données. La partition des dimensions est la suivante : $D_I = \emptyset, D_R = \{\text{CID}\}, D_T = \{Date\}$ et $D_A = \{City, Cust\text{-}Grp, A\text{-}Grp, Product\}$

A partir de la partition de l'ensemble des dimensions, un *item multidimensionnel* e est un m-uplet défini sur les dimensions d'analyse D_A ou à l'aide du symbole $*$. Plus précisément, $e = (d_1, d_2, \ldots, d_m)$ tel que $d_i \in Dom(D_i) \cup \{*\}, \forall D_i \in D_A$ et où $*$ joue le rôle de valeur *joker*.

Par exemple, $(NY, Educ., Middle, A)$ et $(*, *, Young, A)$ sont deux items multidimensionnels par rapport aux quatre dimensions d'analyse de D_A précédemment définies.

La notion d'itemsets et de séquences multidimensionnels découlent naturellement de cette notion d'item multidimensionnel. Ainsi, un *itemset multidimensionnel* $i = \{e_1, \ldots, e_k\}$ est un ensemble non vide d'items multidimensionnels. Par rapport à la définition d'ensemble, deux items *comparables* ne peuvent pas être dans le même itemset. Par exemple, $\{(NY, *, M, B), (*, Educ, Y, C)\}$ est un itemset multidimensionnel alors que $\{(NY, *, M, B), (*, *, *, B)\}$ n'en est pas un puisque $(NY, *, M, B) \subseteq (*, *, *, B)$. Dans ce cas, on dit que $(NY, *, M, B)$ est plus spécifique que $(*, *, *, B)$. On dit aussi qu'un item e est le *plus spécifique* s'il n'existe pas d'item e' tel que $e' \subseteq e$.

Une *séquence multidimensionnelle* $s = \langle i_1, \ldots, i_l \rangle$ est une liste ordonnée d'itemsets multidimensionnels. Par exemple, $s_1 = \langle \{(NY, *, M, A), (*, Educ, Y, B)\}\{(*, *, M, C)\}\rangle$ est une séquence multidimensionnelle.

Chaque bloc défini sur D_R identifie une séquence multidimensionnelle de données. La base exemple de la figure Fig. 1 contient trois blocs différents identifiés par $CID =$ "C_1", $CID =$ "C_2" et $CID =$ "C_3".

Un bloc *supporte* une séquence s si on peut retrouver dans la séquence de données identifiéer par ce bloc tous les items de tous les itemsets de s tout en respectant la relation d'ordre introduite par D_T.

Définition 1 (Séquence et Bloc) *Un bloc de données B_r supporte une séquence $\varsigma = \langle i_1, \ldots, i_l \rangle$ si :*

- $\forall j = 1 \ldots l, \exists d_j \in Dom(D_t), \forall e = (a_{i_1}, \ldots, a_{i_m}) \in i_j, \exists c = (f, r, (x_{i_1}, \ldots, x_{i_m}), d_j) \in B_r$ *avec* $a_i = x_i$ *ou* $a_i = *$
- $d_1 < d_2 < \cdots < d_l$.

Le *support absolu* d'une séquence multidimensionnelle s dans une base de données DB correspond au nombre de blocs de B_{DB,D_R} qui contiennent s. Le *support relatif* correspond au pourcentage de blocs de B_{DB,D_R} qui contiennent s ($\frac{absolute\_support(S)}{|B_{DB,D_R}|}$).

Etant donné un seuil de support fixé Apriori par l'utilisateur, noté σ ($0 < \sigma \leq 1$), une séquence s est *fréquente* sur la base de données DB si $relative\_support(S) \geq \sigma$.

Le but de l'extraction des motifs séquentiels multidimensionnels est de découvrir l'ensemble complet des séquences fréquentes, étant donnés une base de données DB et un seuil de support minimum σ.

Les motifs séquentiels multidimensionnels permettent de mieux décrire les données étudiées. En effet, les corrélations sont extraites sur plusieurs dimensions et une relation d'ordre. Toutefois, les motifs présentent des limites dès lors qu'on se situe dans des contextes où les dimensions ne sont pas seulement symboliques.

Limites des motifs séquentiels multidimensionnels

Les données de productions des entreprises, des administrations, sont souvent agrégées dans un entrepôt de données dans des fins d'analyse. Ainsi, une (ou plusieurs) dimension particulière appelée *mesure*, matérialise le résultat de cette agrégation. Cette dimension est numérique. Elle représente le résultat d'agrégation des données transactionnelles telle que la somme ou le count. La fonction d'agrégation dépend de la sémantique de l'application. Dans cet article, nous considérons l'opérateur d'agrégation sum par défaut.

$$D_1 \times D_2 \times \ldots \times D_n \rightarrow M$$

$$(d_1, d_2, \ldots, d_n) \mapsto m$$

La figure 2 représente un exemple de cube de données résultant de l'agrégation de données transactionnelles issues de bases telles que la figure 1. Puisque les données sont agrégées dans une perspective d'analyse, la notion d'individu (CID) disparaît au profit de groupe d'individus (customer-grp, customer-age, etc.). De plus, une nouvelle dimension apparaît : la mesure. Il est donc nécessaire de faire une nouvelle partition de l'ensemble des dimensions $D \cup M$. Considérons la partition suivante :

- $D_T = \{Date\}$
- $D_R = \{Cust\text{-}Grp\}$
- $D_A = \{City, A\text{-}Grp, Product, Measure\}$

Date	City	Customer Informations		Product	Mesure
1	NY	$Educ.$	$Middle$	A	123
1	NY	$Educ.$	$Middle$	B	234
2	LA	$Educ$	$Middle$	C	120
1	SF	$Prof.$	$Middle$	A	125
2	SF	$Prof.$	$Middle$	C	115
1	DC	$Business$	$Retired$	A	1
1	LA	$Business$	$Retired$	B	24

FIG. 2 – *Datacube*

Puisque la notion d'individu a disparu (CID), nous prenons comme dimension de référence le groupe de consommateur. Cette dimension permet d'identifier 3 blocs comme l'illustre la figure 3. En effet, D_R permet d'identifier les blocs $B_{educ.}, B_{prof.}$ et $B_{business}$. La relation d'ordre reste la même ($D_T = \{Date\}$). La mesure est intégrée dans les dimensions d'analyse. Par rapport aux définitions précédentes, il est très intuitif de traiter cette dimension comme une dimension d'analyse et considérer seulement les cellules qui ont une mesure non vide.

Date	City	Customer Informations		Product	Mesure
1	NY	**Educ.**	$Middle$	A	123
1	NY	**Educ.**	$Middle$	B	234
2	LA	**Educ.**	$Middle$	C	120
1	SF	**Prof.**	$Middle$	A	125
2	SF	**Prof.**	$Middle$	C	115
1	DC	**Business**	$Retired$	A	1
1	LA	**Business**	$Retired$	B	24

FIG. 3 – *Block partition according to* $D_R = \{Cust\text{-}Grp\}$

L'extraction de motifs séquentiels multidimensionnels s'appuie sur une gestion symbolique des données qu'elle traite. Ainsi, étant donnée la partition précédente, l'extraction de motifs séquentiels multidimensionnels a pour objectif de découvrir des corrélations entre la ville, l'âge des consommateurs, les produits vendus et la mesure associée au cours du temps. Cependant, les motifs extraits présentent des limites non négligeables dus à la gestion symbolique de la mesure. En effet, en se basant sur les définitions précédentes, nous pouvons obtenir les situations suivantes :

- Le support absolu de la séquence $\langle\{(*, M, A, 125)\}\rangle$ est 1. En effet, seul le bloc $\mathcal{B}_{Prof.}$ supporte cette séquence. Le bloc $\mathcal{B}_{Educ.}$ contient une séquence relativement similaire $\langle\{(*, M, A, 123)\}\rangle$. Toutefois, la gestion symbolique de la dimension numérique qu'est la mesure fait que les valeurs 123 et 125 sont considérées comme totalement différentes.
- Le support de la séquence $\langle\{(*, *, A, *)\}\rangle$ est 3. Les trois blocs supportent donc la séquence. Plus précisément, les items des séquences de données qui supportent la séquence (l'item) sont $(*, *, A, 123)$ pour $\mathcal{B}_{Educ.}$, $(*, *, A, 125)$ pour $\mathcal{B}_{Prof.}$ et $(*, *, A, 1)$ pour $\mathcal{B}_{Business}$. Nous omettons les valeurs instanciées sur la ville, et l'âge pour mieux faire

ressortir l'observation suivante. $(*, *, A, 125)$ et $(*, *, A, 1)$ ont le *même impact* dans le calcul du support de la séquence $\langle\{(*, *, A, *)\}\rangle$.

Les deux points précédents soulignent les limites d'une gestion symbolique de la mesure dans l'extraction de motifs séquentiels multidimensionnels quand celle-ci est incluse dans les dimensions d'analyse. Il est donc nécessaire de prendre en compte la spécificité de cette dimension : son caractère numérique.

3 La mesure comme valeur numérique : panorama des travaux associés

La présence de valeurs numériques pour des « approches symboliques » est un problème relativement étudié. Ainsi Laurent (2003) propose une architecture basée sur les bases de données multidimensionnelles floues pour générer des résumés flous. Dubois et al. (2003, 2006) s'intéressent à ce problème dans le cadre de l'extraction de règles d'association sur des attributs numériques. Messaoud et al. (2006) utilisent la mesure afin de calculer le support et la confiance des règles d'association recherchées entre les positions des cellules d'un cube de données. Fiot et al. (2005) utilisent la théorie des sous-ensembles flous pour prendre en compte les attributs numériques dans le contexte de la recherche de motifs séquentiels.

Dans Fiot et al. (2007), les auteurs proposent de discrétiser le mesure à l'aide de partitionnement stricts et flous afin de considérer la mesure comme une dimension d'analyse. Intuitivement, une cellule $\langle(a, b, c) : 2\rangle$ d'un cube de données devient $\langle(a, b, c) : \text{peu}\rangle$. La prise en compte de la mesure comme dimension d'analyse présente certaines limites non négligeables :

- Il faut effectuer un prétraitement des données afin de discrétiser cette dimension numérique. Ce pré-traitement peut s'avérer coûteux.
- Une telle approche n'est pas forcément adaptée dans un contexte où des tendances générales sont recherchées. En effet, un cellule ayant une faible mesure peut être vue comme le résultat de l'agrégation de faits non-fréquents. Ainsi, il n'est pas pertinent d'extraire des connaissances issues de l'agrégation de faits non fréquents dans le but de découvrir les tendances générales sur le cube de données. Cette approche peut être particulièrement performante pour des applications qui s'appuient sur la recherche de connaissances rares, anormales ou inattendues.

A notre connaissance, il n'existe pas d'approche qui utilise la mesure pour extraire des motifs séquentiels multidimensionnels. Nous proposons d'utiliser cette dimension numérique particulière pour calculer le support des séquences multidimensionnelles et déterminer ainsi si elles sont fréquentes ou non.

4 La mesure pour calculer le support

Dans la plupart des cas, les valeurs des agrégats d'un cube de données peuvent être vues comme un pré-calcul du support de certaines séquences. En effet, une cellule peut être vue comme une séquence d'un item (défini sur les dimensions d'analyse de la cellule). La mesure de la cellule quantifie l'aptitude de la cellule à supporter l'item. Ainsi, une cellule dont la mesure associée est 100, ne doit pas être considérée de la même façon qu'une cellule qui a une

mesure nettement inférieure. Nous proposons ici de prendre en compte la valeur des agrégats afin de calculer le support des séquences multidimensionnelles.

Il est nécessaire de maintenir l'ordre d'apparition des événements dans la séquence ainsi que l'une des propriétés fondamentales inhérentes à l'extraction de motifs (multidimensionnels ou non) : *l'antimonotonie du support.* Soit un motif p, quel que soit P, un super motif de p, on a :

$$support(p) \geq support(P)$$

Tous les algorithmes d'extraction de motifs se basent sur cette propriété afin de parcourir efficacement l'espace de recherche pour extraire tous les motifs fréquents. Ainsi, ils partent de la séquence vide $\langle\rangle$ et essaient d'extraire des séquences plus longues, soit par un parcours niveau par niveau (APriori, Agrawal et Srikant (1995); Masseglia et al. (1998)), soit en profondeur d'abord (classe d'équivalence Zaki (2001), Pattern-growth Pei et al. (2004)). L'extraction de motifs séquentiels (multidimensionnels) a pour objectif d'établir des corrélations entre des événements suivant leur chronologie d'apparition. Il est ainsi nécessaire de maintenir l'ordre en les éléments d'une séquences.

Pour préserver l'ordre d'apparition des événements dans la séquence ainsi que l'antimonotonie du support, nous utilisons une *t-norme* $\otimes$ qui est une généralisation de la conjonction logique. Une t-norme est un opérateur $[0,1] \times [0,1] \rightarrow [0,1]$ qui est associatif, commutatif, monotone et qui satisfait les conditions s $\alpha \otimes 0 = 0$ et $\alpha \otimes 1 = \alpha$. Les exemples les plus connus de t-norme sont le minimum $(\alpha, \beta) \mapsto \min(\alpha, \beta)$, le produit $(\alpha, \beta) \mapsto \alpha\beta$ et la t-norme de Lukasiewicz $(\alpha, \beta) \mapsto \max(\alpha + \beta - 1, 0)$. Nous utiliserons ici le $\min$ comme t-norme dans les exemples.

Nous utilisons également une *t-conorme* $\oplus$ qui correspond à une disjonction logique. $\oplus$ est un opérateur $[0,1] \times [0,1] \rightarrow [0,1]$ qui est associatif, commutatif, monotone et qui satisfait les conditions $\alpha \oplus 1 = 1$ et $\alpha \oplus 0 = \alpha$. Les exemples les plus connus de t-conorme sont le maximum $(\alpha, \beta) \mapsto \max(\alpha, \beta)$, la somme probabiliste $(\alpha, \beta) \mapsto \alpha + \beta - \alpha\beta$, la somme bornée $(\alpha, \beta) \mapsto \min(\alpha + \beta, 1)$, etc. Nous utiliserons ici le $\max$ comme t-conorme dans nos exemples.

Etant donné qu'une séquence peut apparaître plusieurs fois dans une séquence de données identifiée par un bloc, il est nécessaire d'exhiber la combinaison qui « supporte le mieux » la séquence. Plus précisément, il faut exhiber les cellules qui ont la plus forte valeur de mesure et qui permettent de supporter la séquence.

Pour calculer le support relatif d'une séquence multidimensionnelle, nous avons deux possibilités :

1. L'utilisateur peut considérer que l'importance des blocs doit s'exprimer dans le calcul du support d'une séquence. Ainsi, les blocs ont des poids différents en fonction de leur **effectif ou population**. L'importance d'un bloc intervient dans la valeur du support de la séquence. Par exemple, un bloc important à un impact plus important dans le support d'une séquence qu'un bloc de poids faible.

2. Comme pour les motifs séquentiels classiques où les dimensions de références D_R sont réduites à un singleton représentant l'identifiant d'une séquence de données (*e.g.* l'identifiant du client dans le contexte de l'analyse du panier de la ménagère), les blocs peuvent avoir des impacts égaux dans le support d'une séquence, et ceci, quel que soit leur effectif.

Nous définissons ainsi deux façons de calculer le support relatif d'une séquence dans un cube de données suivant les deux points décrits précédemment.

Micro count prend en compte l'importance de chaque bloc lors du calcul du support d'une séquence. Ainsi, la mesure des cellules d'un bloc qui participent à supporter la séquence est divisée par la mesure totale ($m[cell(*,*,\ldots,*)]$).

Définition 2 (Micro Count) *Soit une g-k-séquence $s = \langle s_1, s_2, \ldots, s_g \rangle$, le support relatif de s dans un cube de données DB avec la technique micro count est égal à :*

Relative support$(s) = \sum_{\mathcal{B}_r \in B_{DB,D_R}} \bigoplus^{\theta_{\mathcal{B}_r}} \bigotimes_{s_i \in s} \bigotimes_{s_{i_j} \in s_i} \frac{(m[\mathcal{B}_r, s_{i_j}])}{m[(*,*,\ldots,*)]}$

Pour chaque séquence apparaissant dans un bloc (tous les items de tous les itemsets doivent être présents en respectant la relation d'ordre), nous prenons la valeur minimale (t-norme) de la valeur de la mesure parmi les cellules supportant les items de la séquence (permet de garantir l'antimonotonie du support).

Puisqu'une séquence peut apparaître plusieurs fois dans la séquence de données pointée par la bloc, il faut considérer la meilleure solution, c'est-à-dire la combinaison la plus prometteuse. C'est pour cela que le support maximum (t-conorme) de cette séquence dans le bloc est retenu.

Macro count vise à calculer le support relatif d'une séquence en considérant que chaque bloc du cube de données doit avoir le même impact dans le support d'une séquence. Ainsi, la mesure des cellules d'un bloc B_r permettant à B_r de supporter la séquence recherchée est divisée par la valeur de mesure associée à B_r ($m[r,*,*,\ldots,*]$).

Définition 3 (Macro Count) *Soit une g-k-séquence $s = \langle s_1, s_2, \ldots, s_g \rangle$, le support relatif de s dans un cube de données DB avec la technique macro count est égale à : Relative support*$(s) = \frac{1}{|B_{DB,D_R}|} \times \sum_{\mathcal{B}_r \in B_{DB,D_R}} \bigoplus^{\theta_{\mathcal{B}_r}} \bigotimes_{s_i \in s} \bigotimes_{s_{i_j} \in s_i} \frac{(m[\mathcal{B}_r, s_{i_j}])}{m[(r,*,\ldots,*)]}$

Comme pour la définition 2, il faut rechercher la meilleure combinaison de cellules (t-conorme) afin que le support de la séquence dans le bloc soit maximal. Pour chaque combinaison, il faut garantir l'antimonotonie du support, la valeur de mesure la plus faible (t-norme) des cellules de la combinaison est retenue.

Le calcul du support des items contenant une ou plusieurs valeurs jokers est assez simple. En effet, puisque nous considérons les mesures associées des cellules qui contiennent au plus un item de la séquence pour un bloc donné afin de calculer le support de la séquence. Ainsi, lorsqu'une valeur joker est présente dans un item de le séquence, il faut récupérer la mesure maximale parmi les cellules qui supportent cet item (exhiber la date où la mesure est maximale).

L'exemple 1 illustre le calcul des supports relatifs de plusieurs séquences en fonction de micro count et de macro count. La mesure considérée dans cet exemple est un dénombrement ($count$) sur des dimensions additives.

Exemple 1 *Soit le cube de données présenté dans la figure Fig. 2. La partition des dimensions est la suivante :*

- $D_T = \{Date\}$
- $D_R = \{Cust\text{-}Grp\}$
- $D_A = \{City, A\text{-}Grp, Product\}$

D_R	D_A			M
Educ.	*	*	*	477
Prof.	*	*	*	240
Business	*	*	*	25
*.	*	*	*	742

TAB. 1 – *Les valeurs des différents blocs et la mesure totale*

B_{Educ}					B_{Prof}					$B_{Business}$				
1	NY	$Middle$	A	123	1	SF	$Middle$	A	125	1	DC	$Retired$	A	1
1	NY	$Middle$	B	234	2	SF	$Middle$	C	115	1	LA	$Retired$	B	24
2	LA	$Middle$	C	120										

TAB. 2 – DB *partitionnée en* 3 *blocs par rapport à* $D_R = \{Cust\text{-}Grp\}$

Le tableau Tab. 1 représente la mesure associée aux différent blocs $\mathcal{B}_r$ *de* B_{DB,D_R} *ainsi que la mesure totale.*

La séquence $\langle\{(*,*,A)\}\rangle$ *est présente dans les* 3 *différents blocs. Le bloc* B_{Educ} *la supporte par l'intermédiaire de la cellule* $\langle(1, NY, Middle, A) : 123\rangle$*, les blocs* $B_{Prof.}$ *et* $B_{Business}$ *supportent également la séquence avec respectivement les cellules* $\langle(1, SF, Middle, A) : 125\rangle$ *et* $\langle(1, LA, Retired, A) : 1\rangle$*. Dans chaque bloc, cette séquence n'apparaît qu'une seule fois, la t-conorme sera ainsi appliquée sur un seul élément.*

Pour MicroCount, on divise la mesure des cellules par la mesure totale : $\frac{125+123+1}{742} = 0,34$.

Pour MacroCount, on divise la mesure de chaque cellule par la mesure associée au bloc contenant la cellule, puis on calcule la moyenne : $\frac{\frac{123}{477}+\frac{125}{240}+\frac{1}{25}}{3} = 0,27$.

Nous remarquons que pour le comptage avec MicroCount, les blocs $B_{Educ.}$*,* $B_{Prof.}$ *et* $B_{Business}$ *ont des influences respectives d'environ* 64%*,* 32% *et* 4% *dans le calcul du support d'une séquence alors qu'avec le comptage MacroCount, ils ont tous des influences égales.*

La séquence $\langle\{(*, Middle, *)\}\rangle$ *est uniquement présente dans les blocs* $B_{Educ.}$ *et* $B_{Prof.}$*. Elle apparaît plusieurs fois dans ces deux blocs. En effet, cette séquence est supportée aux dates* 1 *et* 2 *pour les deux blocs. La t-conorme va nous permettre de retenir la meilleure solution pour chaque bloc. Pour le MicroCount, le support de la séquence* $\langle\{(*, Middle, *)\}\rangle$ *est égal à :* $\frac{\max(357,120)+\max(125,115)}{742} = 0,65$.

La séquence $\langle\{(*,*,A)\},\{(*, Middle, *)\}\rangle$ *est également présente dans les blocs* $B_{Educ.}$ *et* $B_{Prof.}$*. Une seule combinaison est possible pour chacun de ces deux blocs. Par exemple,*

Séquences	MicroCount	MacroCount
$\langle\{(*,*,A)\}\rangle$	0.34	$0,27$
$\langle\{(*, Middle, *)\}\rangle$	$0,65$	$0,42$
$\langle\{(*,*,A)\},\{(*, Middle, *)\}\rangle$	0.32	$0,24$
$\langle\{(*,*,A),(*,*,B)\}\rangle$	$0,17$	$0,10$

TAB. 3 – *Le support relatif de plusieurs séquences pour MicroCount et MacroCount*

les cellules $\langle(1, NY, Middle, A) : 123\rangle$ *et* $\langle(1, LA, Middle, C) : 120\rangle$ *permettent à* $B_{Educ.}$ *de supporter la séquence. La t-norme (min) va nous permettre de garantir l'antimonotonie du support. Ainsi le support de la séquence* $\langle\{(*, *, A)\}, \{(*, Middle, *)\}\rangle$ *doit être inférieur ou égal aux supports des séquences* $\langle\{(*, *, A)\}\rangle$ *et* $\langle\{(*, Middle, *)\}\rangle$*. Ainsi, pour le comptage MicroCount, le support de la séquence* $\langle\{(*, *, A)\}, \{(*, Middle, *)\}\rangle$ *est égal à :* $\frac{\min(123,120)+\min(125,115)}{742} = 0,32$.

Mise en œuvre

Ces nouveaux types de comptage du support d'une séquence multidimensionnelle peut s'appliquer dans n'importe quelle approche d'extraction de motifs séquentiels multidimensionnels. En effet, elle respecte bien l'antimonotonie du support, ce qui permet aux algorithmes d'extraire l'ensemble complet des séquences fréquentes. Toutefois, il est nécessaire d'adapter les algorithmes afin de permettre la recherche de la « meilleure » combinaison retrouvée dans la séquence de données d'un bloc. En effet, dans les autres approches, « la meilleure solution est la première découverte », dès que la séquence est trouvée dans le bloc, le support de la séquence est incrémenté et le calcul du support de la séquence se continue avec l'analyse du bloc suivant. On peut voir ces approches comme évoluant dans un contexte particulier où il n'y a pas de meilleure solution lorsqu'une séquence est supportée plusieurs fois dans un bloc, elles sont toutes équivalentes. Ainsi, chaque fois qu'une séquence est supportée par un bloc, on ajoute 1 au support de la séquence. Dans notre contexte, si un bloc supporte une séquence, on ajoute une valeur comprise dans l'intervalle $]0, 1]$ au support global de la séquence.

Nous avons adapté l'algorithme d'extraction de motifs séquentiels multidimensionnels clos $CMSP$ Plantevit et al. (2008). Cet algorithme permet de parcourir efficacement l'espace de recherche en évitant d'extraire des connaissances redondantes. En effet, $CMSP$ extrait des motifs séquentiels multidimensionnels *clos*. Un motif multidimensionnel est clos s'il n'existe pas de séquence plus spécifique ayant le même support. Les motifs clos offrent ainsi une représentation condensée des connaissances sans perte d'information, et introduisent des propriétés efficaces d'élagage de l'espace de recherche.

$CMSP$ extrait donc les motifs clos en suivant une approche pattern-growth sans gérer d'ensemble de candidats. Ainsi, chaque fois qu'une séquence préfixe est considérée, l'algorithme vérifie si il est possible d'intercaler un item au sein de la séquence (au début, au milieu, ou à la fin) tout en conservant le support de la séquence. Si c'est possible, alors la séquence considérée n'est pas close.

Un mécanisme similaire permet également d'élaguer efficacement l'espace de recherche en évitant d'explorer des séquences préfixes non prometteuses, c'est-à-dire des séquences s dont on est sûr qu'il n'existera pas de séquence close ayant comme préfixe une séquence s.

5 Experimentations

Nous avons mené des expérimentations des méthodes de comptage du support MacroCount et MicroCount appliquées à $CMSP$.Ces expérimentations ont été menées sur des données synthétiques. En effet, nous avons mené des cubes de données synthétiques. Nous considérons cinq dimensions d'analyse.

Ces expérimentations visent à étudier le nombre de motifs séquentiels multidimensionnels clos ou fréquents en fonction du seuil de support minimum ainsi que le temps d'exécution de l'extraction de tels motifs. Exprimer le nombre de motifs clos extraits en comparaison du nombre total de motifs fréquents nous permet de souligner la puissance représentative d'une telle représentation condensée. Les deux méthodes de comptages (MicroCount et MacroCount) sont étudiées.

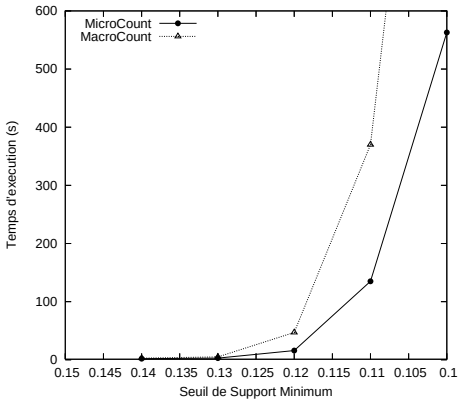
Les courbes Fig. 4(a) et Fig. 4(c) représentent le temps d'exécution de l'extraction des motifs séquentiels multidimensionnels en fonction du seuil de support minimum considéré pour le comptage MicroCount et le comptage MacroCount. Deux jeux de données sont considérés. Le premier (Fig. 4(a)) est partitionné en 25 blocs alors que le second (Fig. 4(c)) est partitionné en 156 blocs. Du fait de l'antimonotonie du support, l'extraction des motifs est plus coûteuse dès que le support diminue. Ceci est inhérent à la problématique d'extraction de motifs. On peut remarquer qu'utiliser le comptage MacroCount est toujours plus coûteux en temps que le comptage MicroCount. En effet, même si les deux courbes suivent le même comportement, le temps d'exécution de l'extraction de motifs avec le comptage MicroCount est toujours plus rapide, dans nos expérimentations, qu'avec le comptage MacroCount.

Les courbes Fig. 4(b) et Fig. 4(d) représentent le nombre de motifs fréquents et le nombre de motifs clos extraits en fonction du seuil de support minimum considéré pour deux jeux de données différents. On remarque que l'utilisation d'une représentation condensée nous permet d'éliminer un nombre important de connaissances redondantes pour les deux méthodes de comptage considérées. Par exemple, 26 motifs séquentiels multidimensionnels clos permettent de représenter 505 motifs séquentiels multidimensionnels pour un support à 0.11 et le comptage MicroCount utilisé dans la Figure Fig. 4(b). Plus précisément, à partir de ces 26 motifs séquentiels clos, il est possible de retrouver toutes les 505 motifs séquentiels multidimensionnels avec leur support exact. Cette puissance représentative est très intéressante, car permet de ne pas présenter à l'utilisateur un ensemble de connaissances redondantes, mais un ensemble plus petit des connaissances qui permettent de retrouver les autres.

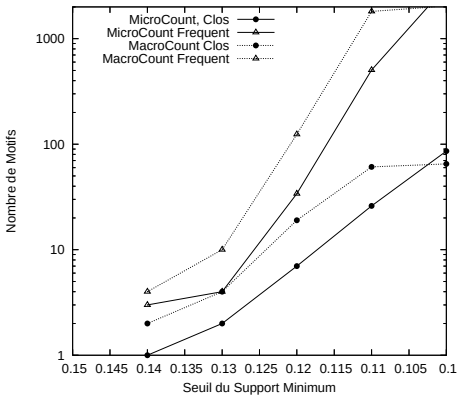
6 Conclusion

Dans cet article, nous avons proposé de directement prendre en compte la mesure dans le calcul du support des motifs séquentiels multidimensionnels. Le calcul du support est une étape clef de l'extraction des motifs puisqu'il permet de définir si un motif est fréquent ou non. Nous avons défini ainsi deux méthodes de comptage s'appuyant sur la mesure des cellules. MicroCount permet ainsi de prendre en compte l'importance du bloc dans le calcul du support d'une séquence alors que MacroCount considère que les blocs doivent avoir un impact égal dans le support d'une séquence. Cette approche permet d'utiliser pleinement les agrégats dans le but d'extraire des connaissances fréquentes par rapport aux nombres de faits présents dans le cube de données. En effet, par ces définitions, nous considérons les mesures des cellules comme des "pré-calculs" potentiels de certains éléments (items) d'une séquences multidimensionnelles.

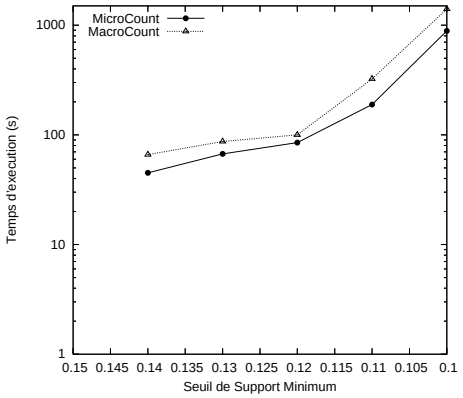
De nombreuses perspectives peuvent être associées à ce travail. Tout d'abord, on peut imaginer utiliser ces méthodes de comptage pour calculer des mesures basées sur le support (confiance, etc.). Il serait ainsi intéressant d'établir la confiance de règles sur des séquences en utilisant la mesure des cellules du cube qui supportent les séquences. Une telle approche pourrait permettre d'extraire des connaissances inattendues ou des exceptions dans un contexte



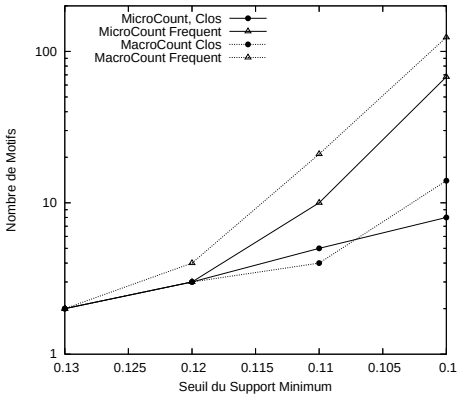
(a) Temps d'exécution en fonction du seuil de support minimum



(b) Nombre de motifs extraits en fonction du seuil de support minimum



(c) Temps d'exécution en fonction du seuil de support minimum



(d) Nombre de motifs extraits en fonction du seuil de support minimum

FIG. 4 – *Expérimentations sur des cubes de données synthétiques*

d'extraction de motifs séquentiels multidimensionnels. Nous devons aussi étudier les cas où les dimensions ne sont pas additives, et notamment voir les répercussions de l'utilisation de la mesure dans un tel contexte (interprétation des résultats, etc.).

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *Proc. 1995 Int. Conf. Data Engineering (ICDE'95)*, pp. 3–14.

Ayres, J., J. Flannick, J. Gehrke, et T. Yiu (2002). Sequential pattern mining using a bitmap representation. In *KDD*, pp. 429–435.

de Amo, S., D. A. Furtado, A. Giacometti, et D. Laurent (2004). An apriori-based approach for first-order temporal pattern mining. In *XIX Simpósio Brasileiro de Bancos de Dados, 18-20 de Outubro, 2004,Brasília, Distrito Federal, Brasil, Anais/Proceedings*, pp. 48–62.

Dubois, D., E. Hüllermeier, et H. Prade (2003). A note on quality measures for fuzzy association rules. In *Proc. of Int. Fuzzy Systems Association World Congress on Fuzzy Sets and Systems*, LNAI 2715, pp. 346–353.

Dubois, D., E. Hüllermeier, et H. Prade (2006). A systematic approach to the assessment of fuzzy association rules. *Data Mining and Knowledge Discovery 13*, 167–192.

Fiot, C., A. Laurent, et M. Teisseire (2005). Motifs séquentiels flous : un peu, beaucoup, passionnément. In *EGC*, pp. 507–518.

Fiot, C., M. Plantevit, et D. Jouve (2007). Quelle partition pour les motifs séquentiels multidimensionnels ? In *Rencontres francophones sur la Logique Floue et ses Applications (LFA'07)*.

Han, J. (1998). Towards on-line analytical mining in large databases. *SIGMOD Record 27*(1), 97–107.

Hsu, J.-L., C.-C. Liu, et A. L. P. Chen (2001). Discovering nontrivial repeating patterns in music data. *IEEE Transactions on Multimedia 3*(3), 311–325.

Laurent, A. (2003). A new approach for the generation of fuzzy summaries based on fuzzy multidimensional databases. *Intell. Data Anal. 7*(2), 155–177.

Lee, C.-H. (2005). An entropy-based approach for generating multi-dimensional sequential patterns. In *PKDD*, Volume 3721 of *Lecture Notes in Computer Science*, pp. 585–592. Springer.

Masseglia, F., F. Cathala, et P. Poncelet (1998). The PSP Approach for Mining Sequential Patterns. In *Proc. of PKDD*, Volume 1510 of *LNCS*, pp. 176–184.

Messaoud, R. B., S. L. Rabaséda, O. Boussaid, et R. Missaoui (2006). Enhanced mining of association rules from data cubes. See Song et Vassiliadis (2006), pp. 11–18.

Pei, J., J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, et M.-C. Hsu (2004). Mining sequential patterns by pattern-growth : The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering 16*(10).

Pei, J., J. Han, B. Mortazavi-Asl, et H. Zhu (2000). Mining access patterns efficiently from web logs. In *PAKDD*, pp. 396–407.

Pinto, H., J. Han, J. Pei, K. Wang, Q. Chen, et U. Dayal (2001). Multi-dimensional sequential pattern mining. In *CIKM*, pp. 81–88.

Plantevit, M., Y. W. Choong, A. Laurent, D. Laurent, et M. Teisseire (2005). M^2SP : Mining sequential patterns among several dimensions. In *PKDD*, pp. 205–216.

Plantevit, M., A. Laurent, et M. Teisseire (2006). Hype : mining hierarchical sequential patterns. See Song et Vassiliadis (2006), pp. 19–26.

Plantevit, M., A. Laurent, et M. Teisseire (2008). Extraction de motifs séquentiels multidimensionnels clos sans gestion densemble de candidats. In *EGC*, pp. 541–546.

Song, I.-Y. et P. Vassiliadis (Eds.) (2006). *DOLAP 2006, ACM 9th International Workshop on Data Warehousing and OLAP, Arlington, Virginia, USA, November 10, 2006, Proceedings*.

ACM.

Srivastava, J., R. Cooley, M. Deshpande, et P.-N. Tan (2000). Web usage mining : Discovery and applications of usage patterns from web data. *SIGKDD Explorations 1*(2), 12–23.

Yang, J., W. Wang, P. S. Yu, et J. Han (2002). Mining long sequential patterns in a noisy environment. In *SIGMOD '02 : Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, New York, NY, USA, pp. 406–417. ACM Press.

Yu, C.-C. et Y.-L. Chen (2005). Mining sequential patterns from multidimensional sequence data. *IEEE Transactions on Knowledge and Data Engineering 17*(1), 136–140.

Zaki, M. J. (2001). Spade : An efficient algorithm for mining frequent sequences. *Machine Learning 42*(1/2), 31–60.

Summary

Data warehouses are now well recognized as the way to store historical data that will be then be available for future queries and analysis. Multidimensional sequential pattern mining aims at discovering correlations among several dimensions through time. Even if multidimensional sequential patterns provide a better view of data by taking data cube specificities into account (e.g. multidimensionality, hierarchies, time), there is no method that propose to take agregate into account in the extraction of such patterns. In this paper, we propose two new definitions of multidimensional sequence support. These definitions are based on the measure of a data cube. Some experiments are reported and emphasize the interest of our proposal.