SALINES : un automate au service de l'extraction de motifs séquentiels multidimensionnels

Yoann PITARCH*, Lionel VINCESLAS**
Anne LAURENT*, Pascal PONCELET*, Jean-Emile SYMPHOR**

*LIRMM - Université Montpellier 2, CNRS
{pitarch,laurent,poncelet}@lirmm.fr
**CEREGMIA, Université des Antilles et de la Guyane, Martinique, France
{lionel.vinceslas,je.symphor}@martinique.univ-ag.fr

Résumé. Les entrepôts de données occupent aujourd'hui une place centrale dans le processus décisionnel. Outre leur consultation, une des finalités des entrepôts est de servir de socle aux techniques de fouilles de données. Malheureusement, les approches existantes exploitent peu les particularités des entrepôts (multidimensionnalité, hiérarchies et données historiques). Parmi ces méthodes, l'extraction de motifs séquentiels multidimensionnels a récemment été étudiée. Nous montrons dans cet article que ces dernières ne tirent pas pleinement profit des hiérarchies et ne découvrent par conséquent qu'une partie seulement des motifs qualitativement intéressants. Nous proposons alors une méthode d'extraction de motifs séquentiels multidimensionnels basée sur un automate et extrayant de nouveaux motifs. Les différentes expérimentations menées sur des jeux de données synthétiques attestent des bonnes performances de notre proposition.

1 Introduction [1]

Initiallement introduites pour faciliter le processus de prises de décisions, les bases de données multidimensionnelles (Codd et al. (1993)) sont de plus en plus utilisées comme support à la fouille de données Han (1998). Dans la mesure où les entrepôts de données stockent des donnés historisées, il apparaît pertinent d'y rechercher des corrélations temporelles. Les motifs séquentiels (Agrawal et Srikant (1995)) sont une des principales techniques utilisées dans cet objectif. Ces motifs, de la forme $\langle(pain, lait)(beurre)(pain, vin)\rangle$, ont été étudiés depuis une dizaine d'années et sont aujourd'hui appliqués dans de nombreux domaines (e.g., profilage de client, détection de fraudes). Cependant, très peu d'approches se sont intéressées à extraire de tels motifs dans un contexte multidimensionnel (Plantevit et al. (2006, 2008)). Par exemple, la découverte du motif $\langle\{(pain, epicerie)(lait, supermarche)\}\{(beurre, epicerie)\}\{(pain, supermarche)(vin, supermarche)\}\rangle$ signifierait que de nombreux clients ont acheté du pain à l'épicerie et dans la même période du lait au supermarché puis du pain au supermarché et enfin du vin au supermarché. Ici, considérer deux dimensions (le produit et le type de magasin) rend les motifs extraits plus intéressants car contenant plus d'informations. De même, il est également possible de prendre en compte les hiérarchies associées aux dimensions (Plantevit et al. (2006, 2008)). Dans HYPE (Plantevit et al. (2006)), il n'est pas possible d'obtenir des motifs où plusieurs niveaux de granularité d'une même dimension apparaissent. Par

1. Ce projet fait parti du projet ANR MIDAS (ANR-07-MDCO-008)

exemple, il n'est pas possible d'obtenir un motif contenant à la fois l'item $(pain, epicerie)$ et $(nourriture, epicerie)$. Dans Plantevit et al. (2008), il est possible d'extraire de tels motifs mais ils ne peuvent être ordonnés que de manière ascendante (motifs divergents) ou de manière descendante (motifs convergents). Aucune spécialisation (resp. généralisation) n'est autorisée si le motif est ascendant (resp.descendant).

Dans cet article, nous proposons une méthode originale, SALINES, exploitant plus efficacement les hiérarchies afin d'extraire des séquences d'items multidimensionnels inédites (de type *"montagne russe"*). Pour cela, nous utilisons une structure d'automate et procédons en deux étapes. Dans un premier temps, un automate des sous-séquences de la base est construit. Ensuite, un parcours de cet automate est réalisé afin d'extraire les motifs multidimensionnels et multiniveaux.

2 Définitions préliminaires

Soit SDB un ensemble de séquences de données multidimensionnelles. Chaque élément appartenant à une séquence est appelé *item* et est décrit sur un ensemble de M dimensions d'analyse noté $\mathcal{D}_A$. A chaque dimension $D_i \in \mathcal{D}_A$ est associé un ensemble de valeurs noté $Dom(D_i)$. Nous supposons que pour chaque dimension D_i, il existe une valeur particulière de $Dom(D_i)$ notée ALL_i et qui signifie *toutes les valeurs de* D_i. Nous supposons qu'il existe une hiérarchie H_i associée à chaque dimension $D_i \in \mathcal{D}_A$. Une hiérarchie H_i est un graphe acyclique orienté où la racine est ALL_i et les nœuds sont les éléments de $Dom(D_i)$. Classiquement, les arcs de ce graphe peuvent être considérés comme des relations *partie de*. Nous notons $H_i^1 : ALL_i > \ldots > H_i^{max_i}$ les max niveaux de précision (granularité) de la hiérarchie H_i. Soit $e \in Dom(D_i)$, nous notons $e \in Dom(H_i^j)$ pour spécifier que l'élément e est décrit sur le niveau de précision H_i^j. Les séquences de SDB sont définies sur les plus fins niveaux de granularité des hiérarchies associées aux dimensions de $\mathcal{D}_A$. Nous supposons que l'utilisateur définit un ordre total sur les combinaisons de niveaux ($c_1 = (H_1^{max_1}, ..., H_M^{max_M}) < ... < c_{fin} = (H_1^1, ..., H_M^1)$) spécifiant ses préférences d'analyse (i.e., quelles sont les combinaisons à considérer en premier dans le processus d'extraction). Le tableau 1 présente les ventes de produits dans différentes villes réalisées par 3 magasins. Chaque item est décrit sur 2 dimensions : *Lieu* et *Produit*. La hiérarchie H_{Lieu} associée à la dimension $Lieu$ est $H_{Lieu} = ALL_{Lieu} > Continent > Pays > Ville$. La hiérarchie $H_{Produit}$ associée à la dimension $Produit$ est $H_{Produit} = ALL_{Produit} > Catégorie > Type$.

S_{ID}	Séquences de données multidimensionnelles
S_1	⟨ *(Lyon,Vin)(Berlin,Bière)(Madrid,Sangria)* ⟩
S_2	⟨ *(Paris,Vin)(Berlin,Bière)(Barcelone,Sangria)* ⟩

TAB. 1 – *Une base de séquences de données multidimensionnelles* SDB

Un item multidimensionnel $a = (d_1, \ldots, d_M)$ est un n-uplet tel que $\forall i = 1, \ldots, M$ on a $d_i \in Dom(D_i)$ et qu'il existe au moins un i tel que $d_i \neq ALL_i$. Par exemple, *(Paris,Vin)* et *(Europe, Alcool)* sont des items multidimensionnels. Un item $a = (d_1, \ldots, d_M)$ est de bas niveau si $\forall i = 1, \ldots, M$, d_i est une feuille de H_i. Une k-séquence multidimensionnelle

$s = \langle e_1, \ldots, e_k \rangle$ est une suite ordonnée de k items multidimensionnels. Un exemple de séquence multidimensionnelle est $\langle$*(Europe, Alcool)(Europe, Coca Cola)*$\rangle$. Une séquence multidimensionnelle est dite de*bas niveau* si tous les items qui la composent sont de *bas niveau*. On dit qu'une séquence (multidimensionnelle) S supporte une séquence $s = \langle e_1, \ldots, e_k \rangle$ si $\forall e_i \in s$, il existe dans S un item e'_i tel que $e' \subseteq e$ et que la relation d'ordre est respectée. Si l'on considère la base de séquences présentée dans le tableau 1, nous avons S_1 qui supporte la séquence $\langle$*(Lyon,Vin)(Europe,Alcool)*$\rangle$. Le support d'une séquence s est le nombre de séquences de SDB qui supportent s. Soit $minSupp$ un paramètre numérique défini par l'utilisateur. On dit qu'une séquence s est fréquente si son support est supérieur ou égal à *minSupp*. La problématique de l'extraction de motifs séquentiels multidimensionnels peut être définie comme la recherche dans une base de séquences multidimensionnelles de toutes les séquences dont le support est supérieur ou égal au paramètre utilisateur *minSupp*.

3 L'approche SALINES

L'automate des sous-séquences de SDB est d'abord construit puis parcouru afin d'extraire les motifs séquentiels fréquents de type *montagne russe*. La construction de l'automate s'inspire de l'algorithme SPAMS (Vinceslas et al. (2009)) qui propose une méthode de construction incrémentale d'un automate afin d'indexer les motifs séquentiels (monodimensionnels) d'un flot de données. Ensuite, l'automate des sous-séquences initial ne permet que la ne reconnaissant que des séquences de bas niveau, un parcours en profondeur de cet automate est réalisé. L'objectif de ce parcours est de rechercher des états à fusionner permettant de reconnaître des séquences fréquentes plus générales.

3.1 Construction de l'automate des sous-séquences

Dans un premier temps, la base SDB est transformée en une séquence $S_{SDB} = \langle \{S_1, e_1\}, \ldots, \{S_n, e_n\} \rangle$ telle que $\forall k, l$ avec $1 \leq k < l \leq n$, si $S_k = S_l$ on a l'item e_k qui précède l'item e_l dans la séquence identifiée par S_k dans SDB. Si l'on considère la base de la table 1, un début de transformation possible est $\langle \{S_1, (Lyon, Vin)\}, \{S_2, (Lyon, Vin)\}, \{S_1, ...\} \rangle$.

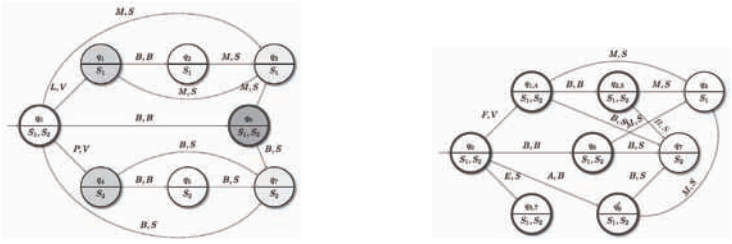
La construction de l'automate des sous-séquences s'appuie sur l'algorithme SPAMS (Vinceslas et al. (2009)) dont nous rappelons brièvement le principe. Chaque état n'est accessible que par un seul item et une valeur de support lui est associée. Cette valeur représente le support de toutes les séquences reconnues par cet état. Par manque de place, il nous est impossible de préciser les détails de la construction de cet automate et invitons le lecteur à se référer à Vinceslas et al. (2009) pour plus de détails. La figure 1(a) présente l'automate des sous-séquences de SDB.

3.2 Extraction des motifs séquentiels multidimensionnels

L'automate généré lors de la première étape ne reconnaît que des séquences de *bas niveau* et n'exploite donc pas les hiérarchies. Etudions la figure 1(a). En considérant $minSupp = 2$, le seul motif séquentiel fréquent reconnu est $\langle$*(Berlin,Biere)*$\rangle$. Pourtant, si l'on observe SDB, nous constatons que la séquence de type *"montagne russe"* $\langle$*(France,Vin)(Berlin,Biere)(Espa-*

gne,Sangria)⟩ est également fréquente. Nous proposons de coupler un parcours en profondeur de l'automate à un mécanisme de fusion d'états pour découvrir de telles séquences.

Nous illustrons les techniques appliquées à chaque état et considérons que l'état q_0 est en cours d'analyse et $minSupp = 2$. Dans un premier temps, nous recherchons l'ensemble des états tels que les séquences reconnues par ces états sont fréquentes. Nous notons cet ensemble d'état E_{fbn}.Ici,$\mathrm{E}_{fbn} = \{q_6\}$ car la séquence ⟨*(Berlin,Bière)*⟩ est fréquente. Ensuite, nous recherchons la plus précise combinaison de niveaux, exceptée c_1, telle qu'il existe au moins un item défini sur cette combinaison pouvant être concaténé aux séquences reconnues par l'état en cours d'analyse afin de produire une nouvelle séquence fréquente. Nous notons I_{fus} l'ensemble de ces items. Ici, si l'on considère que cette combinaison est *(Pays,Produit)* alors $I_{fus} = \{$*(France,Vin),(Espagne,Sangria),(Allmagne,Bière)*$\}$. Pour chacun de ces items it, les états accessibles par une transition labellisée par un item de *bas niveau* dont la généralisation est it sont fusionnés entre eux. Par exemple, les états q_1 et q_4 doivent être fisionnés. Fusionner un état signifie d'abord créer un nouvel état dont les clients associés sont l'union des clients des états à fusionner. Les transitions entrantes du nouvel état sont l'union des transitions entrantes des états à fusionner labellisées par it. Les transitions sortantes du nouvel état sont l'union des transitions sortantes des états à fusionner. Si au cours de cette création de transitions sortantes, différents états sont accessibles par des transitions identiquement labellisées, ces états sont fusionnés (ici, q_2 et q_5 sont fusionnés). Les transitions entre l'état en cours d'analyse et les états à fusionner sont supprimées. Si ces états ne sont plus accessibles par aucun état, alors ils sont supprimés de l'automate. La figure 1(b) présente l'automate obtenu après analyse de q_0.



(a) (b)

FIG. 1 – *Analyse de l'état* q_0

Par manque de place, nous ne pouvons détailler l'analyse de tous les états de l'automate initial. La figure 2 présente néanmoins l'automate obtenu après l'analyse de tous ses états. Nous observons que l'approche proposée permet d'extraire de nouveaux types de motifs (⟨*(France,Vin)(Berlin,Biere)(Espagne,Sangria)*⟩).

4 Expérimentations

Nous présentons ici diverses expérimentations menées sur des jeux de données synthétiques. Nous nous intéressons (1) au nombre de motifs multiniveaux extraits par rapports aux motifs séquentiels fréquents de bas niveaux, (2) au temps d'extraction des motifs et (3) à la mé-

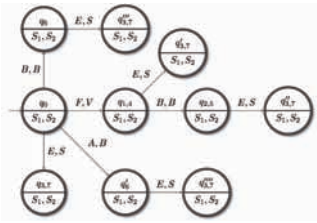
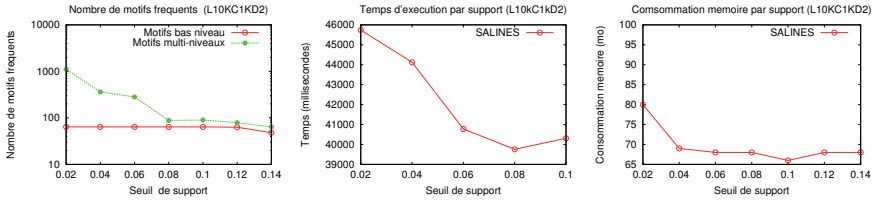


FIG. 2 – *Automate des séquences multiniveaux fréquentes de* SDB

moire vive consommée. Ces expérimentations ont été menées sur un EeePC 1000 he à 1,6Ghz 1Go ram.

Les données multidimensionnelles ont été générées grâce à un générateur de données aléatoires (suivant une distribution aléatoire uniforme) utilisé pour l'évaluation de méthodes de construction de cubes de données. La convention pour nommer les jeux de données est la suivante : L10kC1kD2Lv2 signifie que la taille de S_{SDB} est 10k, que le nombre de clients est 1k et que les données sont décrites sur 2 dimensions observables sur 2 niveaux de granularité (sans le niveau ALL).



(a) Comparaison nombre motifs bas niveaux/nombre motifs multiniveaux (b) Temps d'extraction des motifs multiniveaux (c) Consommation mémoire

FIG. 3 – *Expérimentations sur des données synthétiques (L10kC1kD2Lv2)*

La figure 3 présente les résultats obtenus sur un jeu de données L10kC1kD2Lv2. A travers la figure 3(a), nous constatons que le nombre de motifs multiniveaux est supérieur au nombre de motifs bas niveaux. Par exemple, avec un support faible, nous voyons que, bien entendu, un plus grand nombre d'agrégations s'opèrent et nous obtenons près de 800 séquences supplémentaires sous la forme de "*montagne russe*". Concernant le temps d'exécution de SALINES est très acceptable (3(b)). En effet, avec un support minimum très bas ($minSupp = 0,02$), il est environ de 45 secondes. Le figure 3(c) présente la mémoire consommée en fonction du support minimum. Même avec un support minimum bas, la taille de l'automate reste bornée en mémoire centrale (moins de 80 M0).

De nouvelles expérimentations sont en cours afin d'évaluer pleinement l'impact de la variation de paramètres critiques (hiérarchie, nombre de dimensions). Par exemple, sur un jeu de

données réelles issues de capteurs associées à des pompes industrielles, nous observons que les agrégations effectuées en regroupant par hiérarchie restent bornées en mémoire.

5 Conclusion

Dans cet article, nous apportons une solution originale basée sur un automate afin d'extraire une nouvelle catégorie de motifs séquentiels multidimensionnels : les motifs séquentiels en *"montagne russe"*. Dans un premier temps, l'automate des sous-séquences de bas niveaux est construit puis parcouru afin d'extraire cette nouvelle catégorie de motifs. Au cours de ce parcours, des fusions d'états sont réalisés pour faire émerger des items fréquents généraux. Les premières expérimentations menées confirment que notre approche est applicable sur des bases de données volumineuses. Plusieurs perspectives peuvent être envisagées à la suite de ce travail parmi lesquelles l'extraction de motifs séquentiels multidimensionnels dans un contexte de flots de données.

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pp. 3–14.

Codd, E. F., S. B. Codd, et C. T. Salley (1993). *Providing OLAP (on-line analytical processing) to user-analysts : An IT mandate*. Technical report, EF Codd and Associates, 1993.

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

Masseglia, F., F. Cathala, et P. Poncelet (1998). The PSP approach for mining sequential patterns. In *Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery*, pp. 176–184. Springer-Verlag London, UK.

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*, 1424–1440.

Plantevit, M., A. Laurent, et M. Teisseire (2006). HYPE : mining hierarchical sequential patterns. In *Proceedings of the 9th ACM international workshop on Data warehousing and OLAP*, pp. 19–26. ACM New York, NY, USA.

Plantevit, M., A. Laurent, et M. Teisseire (2008). Up and down : Mining multidimensional sequential patterns using hierarchies. In *Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery*, pp. 156–165. Springer-Verlag Berlin, Heidelberg.

Vinceslas, L., J.-E. Symphor, A. Mancheron, et P. Poncelet (2009). Spams : Une nouvelle approche incrémentale pour l'extraction de motifs séquentiels fréquents dans les data streams. In J.-G. Ganascia et P. Gançarski (Eds.), *EGC*, Volume RNTI-E-15 of *Revue des Nouvelles Technologies de l'Information*, pp. 205–216. Cépaduès-Éditions.

Summary

Nowadays, datawarehouses are well implanted and play a crucial role in the decisional process. Querying datawarehouses is not a sufficient task and data mining techniques adapted to the specificities of this framework (e.g., multidimensionality, hierarchies, historized data) must be proposed. Among theses techniques, multidimensional sequential patterns have been recently studied. In this paper, we show that the existing methods do not take fully into account the hierarchies. Thus, interesting patterns are not extracted. To overcome this drawback, we propose an original multidimensional sequential pattern extraction method based on a automaton which extract new interesting patterns. Experiments conduced on both synthetic datasets show that our approach obtain good performances.