Closed-Pattern : Une contrainte globale pour l'extraction de motifs fréquents fermés

Mehdi Maamar, Christian Bessiere, Patrice Boizumault, Nadjib Lazaar, Yahia Lebbah, Valentin Lemière, Samir Loudni

ClosedPattern : Une contrainte globale pour l'extraction de motifs fréquents fermés *

M. Maamar[1,2] **C. Bessiere**[1] **P. Boizumault**[3] **N. Lazaar**[1] **Y. Lebbah**[2] **V. Lemière**[3] **S. Loudni**[3]

[1] Laboratoire LIRMM, Université de Montpellier, France

[2] Laboratoire LITIO, Université d'Oran, Algérie

[3] Laboratoire Greyc, Université de Caen, France

nom@lirmm.fr, nom.prénom@univ-oran.dz, prénom.nom@unicaen.fr

Résumé

L'extraction de motifs fréquents fermés est un des défis majeurs en fouille de données. Les travaux entrepris récemment en extraction de motifs ont mis en avant l'intérêt d'utiliser les contraintes pour une fouille déclarative. Ces approches se sont montrées très attractives par leurs flexibilité, mais l'utilisation d'un nombre important de contraintes réifiées et de variables auxiliaires posent un sérieux problème quant au traitement des bases de grandes tailles. Dans ce papier, nous présentons une contrainte globale nommée ClosedPattern , qui capture la sémantique particulière des motifs fermés pour résoudre efficacement ce problème, sans faire appel aux contraintes réifiées. Nous proposons un algorithme de filtrage pour la contrainte ClosedPattern, qui maintient la consistance de domaine DC en un temps et espace polynomial.

1 Introduction

La contrainte globale ClosedPattern a pour objectif de palier à deux problèmes majeurs dans la fouille de motifs, à savoir : (i) La rigidité des algorithme classiques [4, 5] pour la prise en compte de nouvelles contraintes. (ii) La lourdeur du modèle réifié proposé dans [1, 2]. En effet, la contrainte globale ClosedPattern permet d'extraire les motifs fréquents fermés d'une base de transactions donnée, sans faire appel aux contraintes réifiées. L'espace de recherche exploré par la contrainte ClosedPattern, est défini uniquement sur les variables de décision représentant les items, contrairement à l'approche déclarative [1], où le modèle nécessite une autre dimension liée aux variables de transactions. La contrainte ClosedPattern repose sur un algorithme de filtrage efficace assurant la consistance de domaine sur chaque nœud de l'arbre de recherche, avec un temps et un espace polynomiale. L'algorithme de filtrage maintient trois règles de filtrage qui permettent d'élaguer toutes valeurs inconsistantes qui ne mènent pas vers un motif fréquent ou fermés. La résolution établit par la contrainte ClosedPattern est sans retour arrière.

*Ce papier est un résumé de l'article publié à CP'16 : "A Global Constraint for Closed Frequent Patterns Mining" [3]

2 ClosedPattern : Encodage et filtrage

Étant donné n items, soit P le motif fréquent fermé recherché, encodé à l'aide des variables booléennes $P_1, ..., P_n$ représentant les items du motif. La contrainte globale ClosedPattern assure à la fois, la propriété de fréquence minimale et la propriété de fermeture. Soit $\mathcal{D}$ la base de transactions et θ le seuil de fréquence minimale. La contrainte $\text{ClosedPattern}_{\mathcal{D},\theta}(P)$ est satisfaite, si et seulement si, $freq_{\mathcal{D}}(P) \geq \theta \wedge P$ forme un motif fermé.

Afin de satisfaire la contrainte ClosedPattern, nous avons proposé trois règles de filtrage pour caractériser l'inconsistance des valeurs 0/1 pour chaque item.

Règle 1 : : Cette règle prend son origine dans la notion de *merging item* proposée dans [6]. Lorsqu'un item i est présent dans toutes les transactions qui couvrent l'instanciation partielle courante, alors, ce genre d'item doit forcement être présent pour former un motif fermé : $0 \notin D(P_i)$.

Règle 2 : : Cette règle est une règle de base, dérivée de la propriété d'anti-monotonie de la fréquence. Si l'ajout d'un item i à l'instanciation partielle courante

forme un motif non fréquent, alors, cet item doit être absent : $1 \notin D(P_i)$.

Règle 3 : Cette règle est originale, elle tire son raisonnement des items absents du motif courant. Si la couverture d'un item i est un sous-ensemble de celle d'un item k. Ainsi, si l'item k est absent $P_k = 0$, alors, l'item i doit également être absent : $1 \notin D(P_i)$.

3 Complexité

Étant donnée une base de transactions $\mathcal{D}$ avec n items et m transactions, et un support minimal θ. L'algorithme de filtrage de la contrainte ClosedPattern maintient la consistance de domaine, ou prouve l'inconsistance dans un temps $O(n^2 \times m)$ avec une complexité en espace en $O(n \times m)$. L'algorithme de filtrage maintient la DC sur des variables booléennes à chaque nœud, ce qui implique que l'arbre de recherche exploré est un arbre binaire entier. Ainsi, la taille de l' arbre est (2 × nombre de nœuds) - 1. Par conséquent, la complexité totale pour extraire l'ensemble des motifs fréquents fermés, noté $\mathcal{C}$, est en $O(\mathcal{C} \times n^2 \times m)$.

4 Expérimentations

Les expérimentations ont été menées sur une série de bases de transactions issues du dépôt FIMI [1]. Les résultats ont montré une domination nette de ClosedPattern par rapport au modèle réifié en terme de nombre de nœuds explorés, de nombre de propagations et de mémoire consommée. Par ailleurs, sur les bases de grandes tailles, le modèle réifié n'est pas capable de charger ces bases en mémoire. En terme de temps de calcul, ClosedPattern domine le modèle réifié, mais reste moins performant que l'algorithme spécialisé LCM.

Dans une seconde étude expérimentale, nous avons étudié une voie prometteuse dans la découverte de motifs ; qui est de poser des contraintes sur un ensemble de k motifs (k-patterns sets). Dans ce contexte, l'intérêt d'un motif est évalué par rapport à un ensemble de motifs. Nous proposons de modéliser et résoudre une instance particulière, où l'objectif est d'extraire les k motifs fermés $\{P_1, ..., P_k\}$ tels que :

(i) $\forall i \in [1, k]$: ClosedPattern(P^i).

(ii) $\forall i, j \in [1, k] : P^i \cap P^j = \emptyset$.

(iii) $\forall i \in [1, k] : lb < |P^i| < ub$.

Cette étude (Fig.1) montre que ClosedPattern a un comportement linéaire en variant k, alors que CP4IM suit une échelle exponentielle et va au-delà du timeout sur les deux bases choisies. L'utilisation de LCM avec un post-traitement sur les k combinaisons possibles des motifs fermés est très couteuse et devient rapidement impraticable.

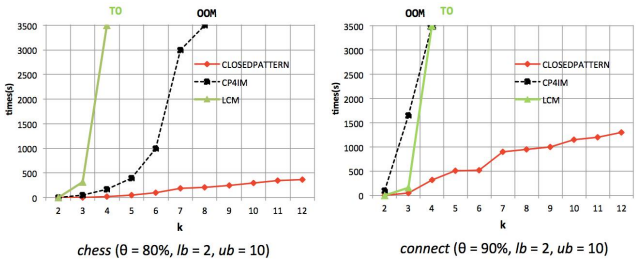

FIGURE 1 – k motifs avec ClosedPattern, CP4IM, LCM

5 Conclusion

Nous avons proposé la contrainte ClosedPattern pour l'extraction de motifs fréquents fermés. Afin de propager efficacement la contrainte, nous avons défini trois règles de filtrage qui assurent la DC. Nous avons conçu un algorithme de filtrage, qui établit la DC avec une complexité cubique en temps et quadratique en espace. Nous avons vu, que ClosedPattern offre un apport pratique sur les bases de grande taille, ce qui est un enjeu majeur pour la communauté de fouille.

Références

[1] L. De Raedt, T. Guns, and S. Nijssen. Constraint programming for itemset mining. In *Proceedings of the 14th ACM SIGKDD*, pages 204–212. ACM, 2008.

[2] T Guns, S Nijssen, and L De Raedt. Itemset mining : A constraint programming perspective. *Artificial Intelligence*, 175(12) :1951–1983, 2011.

[3] N. Lazaar, Y. Lebbah, S. Loudni, M. Maamar, V. Lemière, C. Bessiere, and P. Boizumault. *A Global Constraint for Closed Frequent Pattern Mining*, pages 333–349. CP'16, 2016.

[4] J. Pei, J. Han, and R. Mao. CLOSET : an efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop*, pages 21–30, 2000.

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

[6] J. Wang, J. Han, and J. Pei. CLOSET+ : searching for the best strategies for mining frequent closed itemsets. In *Proceedings of the 9th ACM SIGKDD*, pages 236–245, 2003.

1. http ://fimi.ua.ac.be/data/