Extraction de motifs séquentiels sous contraintes multiples

Nicolas Béchet*, Peggy Cellier**
Thierry Charnois*,*** Bruno Cremilleux*

*GREYC, UMR 6072, CNRS, Univ. de Caen Basse-Normandie,
14032 Caen Cedex, France, {prenom.nom}@unicaen.fr
**IRISA, UMR 6074, INSA Rennes,
35042 Rennes cedex, France, peggy.cellier@irisa.fr
***MoDyCO, CNRS, UMR 7114 Univ. Paris-Ouest Nanterre, 92 001 Nanterre Cedex

1 Introduction

Introduite par Srikant et Agrawal (1996), la fouille de données séquentielles permet de découvrir des corrélations entre des événements selon une relation d'ordre (e.g. le temps). Deux défis majeurs du domaine sont d'une part la définition de méthodes et d'outils permettant d'appréhender de très grands volumes de données et d'autre part la sélection de motifs potentiellement intéressants.

En intégrant des connaissances sous forme d'apriori dans le processus de fouille, l'extraction de motifs sous contraintes (Ng et al. (1998)) apporte des éléments de solution aux défis précédents. En effet, elle contribue à réduire le nombre de motifs en éliminant les motifs a priori inintéressants. Ensuite, elle permet souvent de concevoir des algorithmes plus efficaces en réduisant l'espace de recherche. Pei et al. (2007) ont effectué une étude et proposé un cadre formalisé pour l'extraction de motifs séquentiels sous contraintes. Cependant, cette étude ne prend pas en compte la notion de *contrainte multiple*, une telle contrainte étant définie comme une combinaison de plusieurs contraintes possédant des propriétés antinomiques. Un exemple de contrainte multiple est la contrainte combinant la contrainte de support (qui est anti-monotone), celle d'appartenance d'un item (qui est monotone) et celle de gap entre éléments de la séquence. À notre connaissance, il n'existe pas dans la littérature d'algorithme d'extraction des motifs sous contraintes multiples avec des séquences composées d'itemsets. L'objet de cet article est de traiter ce problème délicat en proposant l'algorithme *PrefixConstraint*.

2 Description de la méthode

PrefixConstraint repose sur la notion de *Pattern Growth* (cf. PrefixSpan, Pei et al. (2001)). Les contraintes multiples y sont traitées en fonction de leurs propriétés de monotonie ou anti-monotonie. Ces dernières ont des propriétés portant sur l'inclusion de motifs qui sont particulièrement adaptées lors de l'utilisation d'algorithmes se fondant sur l'utilisation de bases projetées. En effet tout nouveau motif est soit un motif de taille 1, soit un spécialisation qui

est une extension de son père. Ainsi, les contraintes monotones et anti-monotones sont dans un premier temps prisent en compte lors la génération des motifs de taille 1. Les motifs de taille 1 (i.e. items) ne vérifiant pas les contraintes anti-monotones sont supprimés de la base. Par exemple, si l'utilisateur souhaite obtenir uniquement des motifs sans l'item "10" (contrainte de non appartenance), il est inutile de conserver cet item dans la base. De plus, les séquences de la base ne vérifiant pas les contraintes monotones ne sont pas conservées dans le processus de fouille. Par exemple, si l'utilisateur souhaite extraire des motifs contenant l'item "manger" (contrainte d'appartenance), alors il n'est pas nécessaire de conserver dans la base les séquences ne contenant pas l'item "manger". D'autre part, notre méthode tire profit à la fois des contraintes monotones et anti-monotones. Notons que la contrainte de gap, qui ne vérifie pas une propriété d'anti-monotonie, peut être traitée comme une contrainte anti-monotone (cf. Pei et al. (2007)), car il suffit de vérifier celle-ci entre le préfixe courant et un nouvel itemset. Un prototype à été développé et est actuellement utilisé notamment par des chercheurs en linguistique et en biologie médicale [1].

3 Conclusion

Nous donnons dans cet article les principes de notre algorithme *PrefixConstraint* qui permet d'extraire des motifs séquentiels d'itemsets vérifiant des contraintes multiples. Une perspective de nos travaux est la conception d'un algorithme de fouille combinant certaines techniques de CloSpan et de BIDE pour extraire des motifs séquentiels sous contraintes multiples et produire une représentation condensée de ceux-ci.

Références

Ng, R. T., L. V. Lakshmanan, A. Pang, et J. Hah (1998). Exploratory mining and pruning optimizations of constrained associations rules. pp. 13–24. ACM Press.

Pei, J., J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, et M. Hsu (2001). Prefixspan : Mining sequential patterns by prefix-projected growth. In *ICDE*, pp. 215–224.

Pei, J., J. Han, et W. Wang (2007). Constraint-based sequential pattern mining : the pattern-growth methods. *J. Intell. Inf. Syst. 28*(2), 133–160.

Srikant, R. et R. Agrawal (1996). Mining sequential patterns : Generalizations and performance improvements. In *EDBT*, pp. 3–17.

Summary

We deal with the issue of extracting sequential patterns under multiple constraints. To the best of our knowledge, in the literature there is no algorithm to extract this kind of patterns and we propose a new algorithm to address this problem.

1. https ://sdmc.greyc.fr/