Extraction d'itemsets graduels de haute utilité

Priscile Audrey Fongue Assondji*, Jerry Lonlac**, Norbert Tsopze*

*Département d'informatique, Université de Yaoundé 1, Cameroun
{fongueaudrey0,tsopze.norbert}@gmail.com
** IMT Nord Europe, IMT, Univ. Lille, Centre for Digital Systems, F-59000 Lille, France
jerry.lonlac@imt-nord-europe.fr

1 Contexte et problématique

Les itemsets/motifs graduels présentés sous la forme "plus/moins A, plus/moins B, ..." permettent d'exprimer des covariations entre les différents attributs qui décrivent les données. Ils ont fait l'objet de plusieurs études et de nombreux algorithmes (Lonlac et Nguifo, 2020) ont été proposés pour les extraire à partir des données quantitatives. Le nombre de motifs extraits est souvent élevé et de nombreuses mesures (support, saisonnalité, émergence) ont été proposées pour évaluer ces motifs. Par ailleurs, les motifs de haute utilité permettent d'exprimer d'autres intérêts de l'utilisateur dans la recherche des motifs à travers le concept d'utilité. Ces deux formes de motifs (graduel et de haute utilité) permettent d'exprimer d'une part des covariations et d'autre part l'intérêt de l'utilisateur. Ce papier aborde le problème d'extraction de motifs dits graduels de haute utilité permettant de présenter à l'utilisateur comment une covariation de certains items impactera sur son intérêt dans une base de données quantitative. Nous combinons la mesure d'intérêt utilité avec la gradualité pour extraire les motifs graduels de haute utilité. Pour ce faire, l'approche méthodologique consiste à transformer la base de données originale Δ en une nouvelle base Δ' qui stocke les écarts entre les valeurs du même attribut entre différentes transactions. L'algorithme EFIM (Zida et al., 2017) est modifié pour extraire des motifs graduels de haute utilité de la base Δ'.

2 Fouille d'itemsets graduels de haute utilité

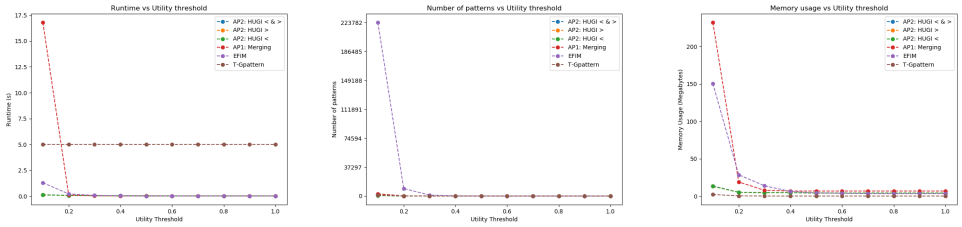
Nous définissons un itemset graduel de haute utilité comme un itemset graduel ayant une utilité supérieure à un seuil fixé par l'utilisateur.

Le principe général de l'extraction d'itemsets graduels de haute utilité suit deux étapes : le codage de la base de données et l'extraction d'itemsets à partir de la base de données encodée.

1. **Codage de la base de données :** Il s'agit de transformer la base de données originale Δ en une base de données Δ' contenant les mêmes attributs que Δ mais dont les occurrences sont les différences entre celles de Δ. Pour une colonne c, les valeurs sont calculées de la manière suivante : ${\Delta'}^c_{kj}$=$\Delta^c_j - \Delta^c_k$. Une valeur positive de Δ'_{kj} représente une augmentation de la quantité lorsqu'on passe de k à j, tandis qu'une valeur négative traduit une diminution.

2. **Extraction d'itemsets graduels de haute utilité :** Deux algorithmes sont proposés à cet effet :(1) diviser la base Δ' en une base des items ayant une variation positive (Δ'_+) et une base des items à variation négative (Δ'_-), puis extraire les itemsets séparemment, puis fusionner les résultats (*HUGI-Merging*) ; (2) extraire directement des itemsets graduels de haute utilité à partir de la base Δ' *(HUGI)* .

3 Résultats expérimentaux



(a) Temps vs minUtil (b) #Motifs vs minUtil (c) #Mémoire vs minUtil

FIG. 1 – Evaluation comparative de HUGI sur les données des transactions commerciales

La figure 1 présente les résultats de l'expérimentation faite sur un ensemble de transactions commerciales (100 items ,418 transactions). Elle montre que HUGI-Merging prend plus de temps et d'espace mémoire pour un seuil inférieur 0.1 dû à la fusion des résultats. A partir du seuil 0.2, le nombre de motifs, le temps et l'espace mémoire diminue et devient quasi constant dû à la présence de plusieurs valeurs égales dans le dataset. Une analyse plus fine montre que HUGI génère moins de motifs que EFIM et T-Gpattern (Lonlac et Nguifo, 2020) comme il élimine les motifs graduels non utiles.

4 Conclusion

Ce papier [1] explore le problème de fouille d'itemsets graduels de haute utilité et propose deux algorithmes qui exploitent l'algorithme *EFIM* pour extraire efficacement de tels itemsets avec une utilité supérieure à un seuil prédéfini. Les expérimentations montrent que l'approche est efficace et extrait moins de motifs que les algorithmes EFIM et T-Gpattern.

Références

Lonlac, J. et E. M. Nguifo (2020). A novel algorithm for searching frequent gradual patterns from an ordered data set. *Intell. Data Anal. 24*(5), 1029–1042.

Zida, S., P. Fournier-Viger, J. C.-W. Lin, C.-W. Wu, et V. S. Tseng (2017). Efim : A fast and memory efficient algorithm for high-utility itemset mining. *Knowl. Inf. Syst. 51*(2), 595–625.

1. Ce travail a été partiellement soutenu par le CNRS à travers le projet AAP-Afrique FDMI-AMG.