Une nouvelle méthode de partitionnement de séquences avec motifs interprétables.

Sébastien Amoury*, Karell Bertet*, Damien Mondou*

* La Rochelle Université - 23 Avenue Albert Einstein, 17000 La Rochelle, France
prenom.nom@univ-lr.fr

1 Contexte

À l'ère du « tout numérique », la génération et l'exploitation de données constituent un défi d'envergure pour la recherche et l'innovation. Le regroupement de données, ou clustering, repose sur de nombreuses approches (Ghosal et al., 2020), mais les solutions traditionnelles ne parviennent à partitionner les données complexes, comme les séquences, qu'après un prétraitement consistant à effectuer un plongement numérique des données. Cette transformation rend difficile l'explication des partitions obtenues. Dans cet article, nous proposons une nouvelle méthode de partitionnement de données séquences à partir d'une hiérarchie de concepts, calculée par l'algorithme NEXTPRIORITYCONCEPT (Demko et al., 2020), issu de l'Analyse Formelle de Concept, avec motifs interprétables.

2 Travaux entrepris et résultats

L'explicabilité est un enjeu de taille lors de traitements automatique des données. Notre cas d'étude porte sur un Serious Game qui s'est déroulé en 2017 (Mondou, 2024). Les traces des joueurs peuvent être assimilées à des séquences d'activités et les comportements communs des joueurs peuvent être observés en calculant un treillis de concepts de leur séquences d'activités (Amoury et al., 2025). Nous proposons une solution automatisée permettant de partitionner les joueurs en sélectionnant des concepts à partir de la hiérarchie du treillis. Soit $(A, \delta(A))$ un concept où A représente la partie objet du concept et $\delta(A)$ un ensemble de prédicats décrivant le motif de A. L'ensemble des concepts dotés d'une relation de spécialisation/généralisation $(A_1, \delta(A_1)) \leq (A_2, \delta(A_2)) \iff A_1 \subseteq A_2 (\iff \delta(A_2) \sqsubseteq \delta(A_1))$ forment un ensemble partiellement ordonné appelé treillis de concepts. Pour des données de type séquences, plusieurs descriptions sont possibles afin de calculer un treillis de concepts, notamment la description par préfixe maximal commun, qui permet d'obtenir un prédicat de la forme "sequence starts with", afin de grouper les joueurs possédant le même préfixe. Un concept $(A, \delta(A))$ correspond donc à un groupe de joueurs A décrits par leur préfixe commun maximal $\delta(A)$. Notre algorithme va extraire un partitionnement sous forme d'une liste de cluster C et d'une liste de motifs M de chaque cluster, pour un seuil support donné en paramètre. Il parcourt le treillis de concepts en partant du concept top, qui contient tous les objets, puis pour chacun des prédécesseurs $(A', \delta(A'))$, triés par ordre décroissant de support, d'un concept courant $(A, \delta(A))$,

va évaluer si ce $(A', \delta(A'))$ respecte le seuil support. S'il le respecte, ce $(A', \delta(A'))$ est ajouté à une liste d'exploration. Sinon si un des prédécesseurs a respecté le seuil pour le concept courant, A' est ajouté à la liste C et $\delta(A')$ est ajouté à la liste M. Sinon A et $\delta(A))$ sont respectivement ajoutés à C et M. Puis, un nouveau $(A, \delta(A))$ est sélectionné dans la liste d'exploration et ce processus se poursuit jusqu'à ce que la liste d'exploration soit vide. Pour finir, un dernier groupe est ajouté à C contenant tous les joueurs restants et le motif associé, ajoutée à M est celui du concept top. Le support étant strictement décroissant, l'algorithme n'a pas besoin de parcourir l'ensemble du treillis afin de sélectionner les concepts. En reprenant l'expérimentation 2 de Amoury et al. (2025), nous avons pu calculer un treillis de concepts des séquences d'activités des 41 joueurs de notre cas d'étude. En utilisant notre algorithme avec en entrée le treillis obtenu précédemment et un seuil support de 3/4, nous obtenons une liste C de 4 groupes de joueurs décrits dans le tableau 1, dont les motifs descriptifs permettent d'en déduire une sémantique.

Partitions formées	Sémantique déduite par le prédicat
$C_1 = \{6, 14, 41, 32, 29, 19\}$	Ont répondu "non" au robot pour jouer au jeu la première fois.
$C_2 = \{31, 33\}$	N'ont pas répondu au robot pour jouer au jeu.
$C_3 = \{18, 36, 37, 13, 22, 35, 11, 10, 3, 23, 2, 5, 38,$	Ont répondu "oui" pour jouer au jeu
$17, 7, 24, 8, 30, 21, 40, 25, 9, 15, 12, 16, 28, 27, 26, 20, 39\}$	et ont entendu la première question.
$C_4 = \{34, 1, 4\}$	Se sont fait scanner par le robot une première fois

TAB. 1 – *Partitionnement obtenu avec notre algorithme pour un seuil de 3/4.*

3 Conclusion

Le travail d'analyse de données massive peut-être facilité lorsque les regroupements obtenus sont explicables par un motif. Mais, lorsque les représentations graphiques comme les treillis sont trop volumineuses, il peut être difficile d'extraire efficacement des groupes. Notre proposition permet d'automatiser ce processus en prenant un treillis de concepts en entrée et un seuil support afin de renvoyer des groupes et leurs motif associé, en partitionnant les données, permettant une compréhension aisée.

Références

Amoury, S., K. Bertet, et D. Mondou (2025). Clustering of serious game traces using formal concept analysis. In *Intelligent Data Engineering and Automated Learning – IDEAL 2024*, Cham, pp. 287–299. Springer Nature Switzerland.

Mondou, D. (2024). Temporal automata for robotic scenario modeling with cit framework. *Multimedia Tools and Applications*, 1–26.

Ghosal, A., A. Nandy, A. K. Das, S. Goswami, et M. Panday (2020). A short review on different clustering techniques and their applications. *Emerging Technology in Modelling and Graphics : Proceedings of IEM Graph 2018*, 69–83.

Demko, Ch., K. Bertet, C. Faucher, J.-F. Viaud, et S. O. Kuznetsov (2020). Nextpriorityconcept : A new and generic algorithm computing concepts from complex and heterogeneous data. *Theoretical Computer Science 845*, 1–20.