Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Révéler des communautés latentes à partir des patterns de clickstream : un cadre d'analyse des plateformes web

Mohsine AABID[1, 2] Patrice Bellot[1] Simon Dumas-Primbault[2]
(1) Laboratoire d'informatique et des systèmes de Marseille, France
(2) OpenEditionLab, France
mohsine.aabid@lis-lab.fr, patrice.bellot@lis-lab.fr,
simon.dumas-primbault@openedition.org

RÉSUMÉ

Comprendre les communautés d'utilisateurs constitue un enjeu important pour l'analyse et l'amélioration des plateformes web. Cet article propose un cadre méthodologique fondé sur les données pour identifier des communautés latentes d'intérêt à partir de patterns de clickstream, sans recourir à des données annotées ni à un profilage préalable des utilisateurs. La méthode consiste à construire des signatures comportementales à partir des clics observés au niveau des sessions, en fonction des différents espaces de navigation sur le web, puis à appliquer des techniques de clustering non supervisé afin de mettre en évidence des segments d'utilisateurs distincts. Nous appliquons ce cadre à OpenEdition, une bibliothèque numérique de grande ampleur, en analysant plusieurs millions de sessions. Des métadonnées complémentaires sont ensuite mobilisées pour interpréter les clusters et évaluer leur interprétabilité ainsi que la prédictibilité de l'appartenance aux clusters à partir de ces métadonnées. Bien que démontré sur une plateforme académique, ce cadre peut être étendu à d'autres environnements web multi-sites.

ABSTRACT

Uncovering Latent Communities from Clickstream Patterns : A Framework for Web Platform Analysis

Understanding user communities is important for analyzing and improving web platforms. This paper proposes a data-driven framework to identify latent communities of interest from clickstream patterns without requiring labeled data or prior user profiling. The method builds behavioral signatures from session-level clicks across different areas of web navigation and applies unsupervised clustering to reveal distinct user segments. We apply this framework to OpenEdition, a large digital library, analyzing millions of user sessions. Complementary metadata are then used to interpret the resulting clusters and to assess their interpretability as well as the predictability of cluster membership from available metadata through a classification approach. This evaluation helps identify which metadata attributes contribute most to distinguishing user segments and how they relate to observed behavioral patterns. Although demonstrated on a scholarly platform, the framework can be extended to other multi-site web environments where understanding behavioral diversity can support analysis, personalization, and service design.

MOTS-CLÉS : Analyse de logs, Bibliothèque numérique, Modélisation utilisateur.

KEYWORDS: Log Analysis, Digital Library, User Modeling.

Actes de la 21e Conférence en Recherche d'Information et Applications.
Volume 1 : articles scientifiques originaux, pages 145–155.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

Les plateformes web hébergeant des contenus divers font face à un défi fondamental : comprendre l'hétérogénéité de leur base d'utilisateurs. Si les métriques agrégées permettent d'observer les tendances globales du trafic, elles masquent la richesse de la diversité comportementale qui structure l'engagement sur ces plateformes. Les utilisateurs arrivent avec des intentions différentes, naviguent selon des stratégies variées et interagissent avec des contenus distincts pourtant, la plupart des plateformes ne disposent pas de méthodes scalables permettant de mettre en évidence ces communautés latentes sans recourir à des enquêtes coûteuses ou à des formes de suivi intrusives.
Une approche prometteuse repose sur l'analyse des données de *clickstream*. Chaque session utilisateur laisse des traces comportementales (schémas de navigation, sélection de contenus et formes d'engagement) qui encodent implicitement les intérêts et les intentions des utilisateurs. Toutefois, transformer ces séquences brutes en segments d'utilisateurs interprétables reste une tâche complexe, en particulier en l'absence de données annotées et dans des populations d'utilisateurs hétérogènes et évolutives.
Pour répondre à cette problématique, nous proposons un cadre méthodologique fondé sur les données permettant d'identifier des communautés latentes d'intérêt à partir de patterns de *clickstream* observés au niveau des sessions. L'approche fonctionne sans profilage préalable des utilisateurs, sans données annotées et sans intervention explicite. Elle consiste à construire des signatures comportementales à partir de la distribution des clics dans différents espaces de navigation sur le web, puis à appliquer des techniques de *clustering* non supervisé afin d'identifier des segments d'utilisateurs cohérents. Ces clusters sont ensuite interprétés à l'aide de métadonnées contextuelles, ce qui permet de traduire des motifs comportementaux abstraits en profils de communautés significatifs.
Nous appliquons ce cadre à OpenEdition, une bibliothèque numérique à grande échelle dédiée aux sciences humaines et sociales, qui propose quatre plateformes intégrées (revues, livres, carnets de recherche et annonces d'événements) et accueille des millions d'utilisateurs à l'échelle internationale. Cet environnement constitue un terrain d'expérimentation riche, combinant une forte diversité de contenus, des ressources multilingues, des profils d'utilisateurs hétérogènes (chercheurs, étudiants, professionnels) ainsi que des métadonnées détaillées facilitant l'interprétation des comportements.
Cette étude s'articule autour de la question de recherche suivante : Une analyse non supervisée des patterns de clics peut-elle révéler des communautés d'intérêt pertinentes, et comment ces communautés peuvent-elles être caractérisées à l'aide de métadonnées externes ? Nous faisons l'hypothèse que les distributions de clics entre différents espaces de navigation fonctionnent comme des signatures comportementales. Une fois regroupées par *clustering*, ces signatures peuvent révéler des segments d'utilisateurs latents distingués par leur orientation disciplinaire, leur origine géographique, leurs préférences linguistiques et leurs modes d'accès à la plateforme. Nos principales contributions sont les suivantes :

1. Un cadre méthodologique généralisable pour détecter des communautés d'intérêt à partir de données de *clickstream*, applicable à toute plateforme web multi-sites.

2. Une méthode simple et efficace pour l'identification des sessions et la détection des *bots*.

3. Une approche méthodologique pour caractériser des clusters comportementaux à l'aide de diverses sources de métadonnées externes.

2 Revue de littérature

Les revues de littérature consacrées à l'analyse des logs dans les bibliothèques numériques ont mis en évidence à la fois les avantages et les limites de cette approche, tout en identifiant plusieurs pistes importantes pour de futures recherches (Agosti *et al.*, 2012; Jamali *et al.*, 2005). Les études quantitatives dans ce domaine ont mobilisé les fichiers de logs pour analyser le comportement des utilisateurs selon différentes perspectives. Certains travaux se sont concentrés sur la production de statistiques descriptives d'usage (Liang & Leng, 2020; Fu *et al.*, 2021), tandis que d'autres ont combiné ces analyses avec des méthodes qualitatives afin de révéler ou d'interpréter des motifs comportementaux que les logs seuls ne permettent pas d'identifier (Alokluk & Al-Amri, 2021; Nicholas *et al.*, 2008).
Des approches plus avancées ont appliqué des techniques de *data mining* afin de dépasser les analyses descriptives de base (Trabelsi *et al.*, 2021; Kovacevic *et al.*, 2010), intégré des métadonnées pour obtenir des éclairages supplémentaires sur les usages (Bogaard, 2018; Liang & Leng, 2020), ou encore développé des modèles d'utilisateurs visant à améliorer la personnalisation des services (Zerhoudi *et al.*, 2022). Par ailleurs, certains travaux se sont intéressés aux communautés d'utilisateurs, soit en étudiant des communautés déjà établies (Fu *et al.*, 2021), soit en les identifiant à partir de termes de recherche partagés (Papatheodorou *et al.*, 2003).
Cependant, à notre connaissance, aucun travail antérieur n'a simultanément extrait et caractérisé des communautés sur la base de leurs motifs comportementaux au sein de bibliothèques numériques. De plus, les études existantes n'ont pas mobilisé les métadonnées comme principal cadre analytique pour expliquer et interpréter ces communautés. Ce travail propose un cadre méthodologique permettant d'extraire des communautés définies par des patterns d'interaction partagés, observés à travers les motifs de clics au niveau des sessions et les niveaux d'engagement puis de les caractériser en analysant les métadonnées associées (telles que la discipline des ressources consultées ou la localisation géographique des utilisateurs). Cette approche vise à apporter des éclairages sur les comportements et les pratiques de recherche d'information des utilisateurs, afin de mieux informer et améliorer les services proposés par les bibliothèques numériques.

3 Étapes de la méthodologie

Notre approche méthodologique se décline en plusieurs étapes séquentielles visant à identifier des communautés d'intérêt à partir des logs d'utilisation de la plateforme. Pour construire des représentations vectorielles pertinentes des sessions utilisateurs, nous nous inspirons du paradigme des systèmes de recommandation, et plus particulièrement des techniques de filtrage collaboratif. Le processus complet comprend : (1) l'identification des sessions, (2) le filtrage des sessions automatisées, (3) la construction et la réduction de la matrice d'interaction session-ressource, et enfin (4) le partitionnement (clustering) suivi de l'évaluation de l'alignement des métadonnées.

Avant de détailler ces étapes, il convient de justifier le choix de la factorisation de matrice comme technique centrale de réduction de dimension. Bien que des méthodes plus sophistiquées issues de l'apprentissage profond (deep learning) soient largement documentées notamment les architectures récurrentes comme GRU4Rec (Hidasi *et al.*, 2016), les modèles d'attention comme SASRec (Kang & McAuley, 2018), ou les approches par graphes comme SR-GNN (Wu *et al.*, 2019) nous avons privilégié la factorisation de matrice pour sa sobriété calculatoire et sa grande interprétabilité. Contrairement aux modèles de type "boîte noire", la factorisation permet de lier directement les composantes

latentes aux ressources consommées, offrant une robustesse accrue face à la parcimonie (sparsity) extrême des logs web tout en facilitant la validation sémantique des communautés. Il est toutefois important de noter que notre architecture est modulaire : elle permettrait tout à fait l'intégration de ces méthodes plus complexes pour générer les représentations de sessions, sans modifier la logique globale de notre cadre d'évaluation.

3.1 Identification des sessions

Pour analyser les préférences des utilisateurs, nous devons d'abord identifier les sessions à partir des fichiers logs bruts, afin d'extraire les ressources co-occurrentes au sein d'une même session utilisateur. Nous nous appuyons initialement sur la méthodologie décrite par (Halfaker *et al.*, 2015) pour calculer les intervalles de temps entre les requêtes. Toutefois, nous étendons cette approche en introduisant un seuil probabiliste flexible (*soft threshold*) pour l'assignation des sessions.

Cette extension est motivée par deux observations clés : (1) les intervalles de temps entre les requêtes suivent une distribution à traîne longue (*long-tail distribution*), ce qui rend difficile l'identification d'un point de coupure net ; et (2) la méthodologie de référence ne parvient pas à capturer les sessions longues (caractérisées par des *referrers* correspondant aux URL des requêtes précédentes). Notre approche calcule la probabilité qu'une requête r appartienne à la même session s comme suit :

$$p(r \in s \mid \text{ref}, \Delta t) = \min\left(1, \frac{1 + \text{ref}}{1 + \alpha(\tau - \Delta t)^2}\right)$$

Où Δt est l'intervalle de temps entre les requêtes, τ est le seuil temporel, *ref* est un indicateur binaire égal à 1 si le *referrer* correspond à l'URL de la requête précédente (0 sinon), et α est un paramètre de pénalité contrôlant le taux de décroissance de la fonction.

Nous avons privilégié cette fonction rationnelle à décroissance polynomiale (de type Cauchy) au détriment d'une fonction exponentielle classique. Ce choix permet d'éviter une convergence trop abrupte vers zéro, offrant ainsi une « traîne » plus permissive qui maintient une probabilité d'assignation non négligeable pour des écarts temporels importants, tant que la continuité est suggérée par le *referrer*.

Nous déterminons un seuil temporel approprié en nous basant sur la distribution des intervalles. Nous effectuons ensuite plusieurs essais avec différentes valeurs de α. Nous sélectionnons la valeur de α en utilisant une approche inspirée de la méthode du coude (elbow method), en identifiant le point où l'augmentation du nombre de sessions converge vers un plateau. Ce choix reflète la nécessité d'équilibrer la cohésion des sessions (s'assurer que les requêtes logiquement liées restent groupées avec leur distinction), tout en garantissant la stabilité des résultats en l'absence de données de référence (*ground-truth labels*).

3.2 Filtrage des sessions automatisées

Avant d'appliquer nos critères comportementaux, nous procédons à une première phase de nettoyage par filtrage basé sur les signatures (*User-Agent filtering*). Cette étape permet d'éliminer les robots d'indexation et les *crawlers* connus (tels que Googlebot ou Bingbot) en identifiant les chaînes de caractères spécifiques dans l'en-tête *User-Agent* des requêtes HTTP.

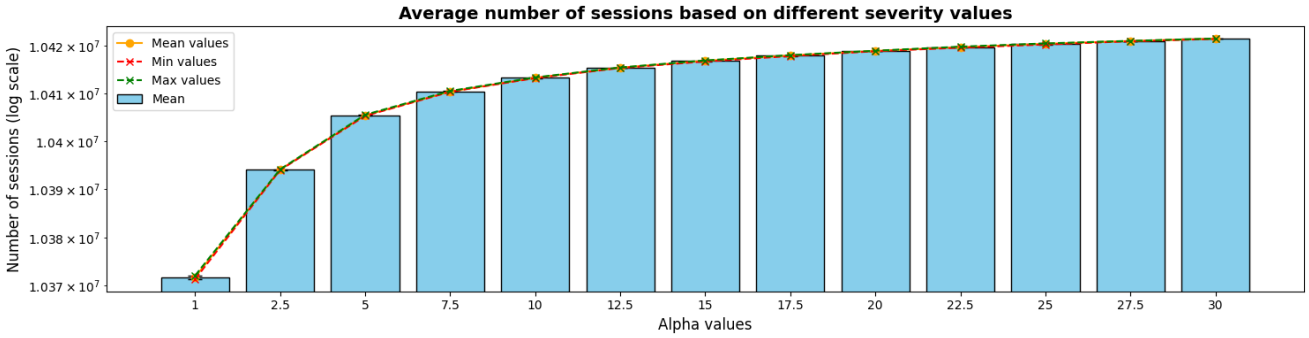

FIGURE 1 – Nombre de sessions en fonction du paramètre α

Toutefois, cette méthode étant insuffisante pour détecter les scripts personnalisés ou les bots simulant des navigateurs standards, nous appliquons ensuite trois heuristiques basées sur des caractéristiques établies dans la littérature (Iliou *et al.*, 2019; Jagat *et al.*, 2024) : le volume de requêtes, la vitesse de navigation et la régularité temporelle. Nous supprimons ainsi les sessions répondant à l'un des critères suivants :

1. **Vitesse de navigation excessive :** Nous fixons un seuil empirique de 2 requêtes par seconde. Nous considérons qu'il est hautement improbable qu'un utilisateur humain puisse sélectionner, cliquer et charger des ressources à une fréquence supérieure sur une période prolongée.
2. **Régularité temporelle parfaite :** Nous écartons les sessions dont l'écart-type des intervalles entre requêtes est strictement égal à zéro ($\sigma = 0$). Une telle régularité est le signe distinctif d'un processus automatisé (script cadencé), le comportement humain étant intrinsèquement sujet à des variations de rythme.
3. **Volume critique de requêtes :** Bien que la distribution du nombre de requêtes par session suive une loi exponentielle (une grande majorité de sessions courtes), l'inspection visuelle de la distribution en échelle logarithmique nous a permis d'identifier un seuil de coupure. Nous filtrons ainsi les sessions dépassant 100 requêtes, limite au-delà de laquelle le comportement diverge des schémas de navigation humaine observés.

3.3 Construction et réduction de la matrice

Nous construisons une matrice de clics M où les lignes représentent les sessions et les colonnes les ressources consultées. Afin de maîtriser les coûts computationnels et d'évaluer la stabilité des clusters, nous échantillonnons les sessions par période temporelle (par exemple par semaine) plutôt que de manière aléatoire, ce qui permet de préserver la distribution temporelle des groupes d'intérêt.

Étant donné que le nombre de ressources est généralement très élevé et que la matrice résultante est particulièrement parcimonieuse (*sparse*), nous appliquons un pré-traitement. Nous supprimons les colonnes correspondant aux ressources rarement visitées, car une ressource consultée dans très peu de sessions n'apporte que peu d'information et introduit du bruit dans le signal global. Ce filtrage est contrôlé par un paramètre $\beta \in (0, 1)$ définissant un seuil minimal de présence proportionnel au nombre total de sessions. Les ressources apparaissant dans une proportion de sessions inférieure à ce seuil sont supprimées. Après l'application de ce filtrage, les sessions devenues vides (lignes nulles) sont également écartées.

Pour obtenir une représentation dense et exploitable des sessions, nous appliquons ensuite une factorisation de matrice non-négative (*Non-negative Matrix Factorization*, NMF) sur la matrice filtrée. Cette méthode approxime la matrice M par le produit de deux matrices de rang réduit

$$M \approx SI,$$

où $S \in R_+^{n \times k}$ et $I \in R_+^{k \times m}$. La matrice S fournit une représentation latente de dimension k pour chaque session. Le choix de la NMF est justifié par le fait qu'elle réduit la dimensionnalité de la matrice de clics en décomposant les sessions en une combinaison additive de signatures de navigation latentes.

Le choix final des hyperparamètres repose sur une recherche par grille (*grid search*) explorant l'interaction entre le paramètre de filtrage β et le nombre de composantes latentes k. Nous retenons la configuration la plus parcimonieuse permettant d'obtenir des représentations stables tout en maximisant la séparabilité des clusters dans l'espace latent.

3.4 Évaluation du partitionnement et alignement des métadonnées

Pour évaluer la qualité des clusters identifiés, nous mobilisons d'abord des métriques classiques de validation interne du partitionnement, notamment l'indice de Davies–Bouldin et le coefficient de Silhouette. Ces mesures permettent d'estimer respectivement la séparation entre clusters et leur cohésion interne dans l'espace latent appris à partir des comportements de navigation.

Cependant, ces métriques structurelles ne suffisent pas à elles seules à déterminer si les clusters obtenus correspondent à des groupes interprétables du point de vue des usages réels de la plateforme. En l'absence de vérité-terrain (ground truth) sur les communautés d'utilisateurs, nous adoptons une stratégie d'évaluation indirecte fondée sur les métadonnées disponibles.

Il est important de souligner que ces métadonnées ne sont volontairement pas intégrées dans la phase de clustering. Notre objectif est de faire émerger des structures comportementales latentes uniquement à partir des patterns de navigation observés dans les logs, sans imposer a priori des catégories éditoriales, géographiques ou thématiques déjà connues. Intégrer ces informations dès l'étape de partitionnement risquerait d'orienter artificiellement les regroupements vers des structures préexistantes plutôt que de révéler des communautés émergentes fondées sur les comportements effectifs de consultation.

Les métadonnées sont donc utilisées dans un second temps comme outil d'interprétation et d'analyse de l'alignement informationnel des clusters. Nous reformulons cette analyse comme une tâche supervisée auxiliaire : un modèle de classification est entraîné pour prédire l'appartenance à un cluster à partir des seules métadonnées.

Cette étape ne constitue pas une validation supervisée du clustering, mais une mesure de la quantité d'information descriptive contenue dans les métadonnées relativement à la structure comportementale découverte de manière non supervisée. Une forte capacité prédictive suggère que certains comportements de navigation sont associés à des caractéristiques descriptives identifiables, tandis qu'une faible capacité prédictive indique au contraire que les clusters capturent des dimensions comportementales plus difficiles à expliquer à partir des seules métadonnées disponibles.

Nous analysons également l'importance des variables afin d'identifier quelles dimensions descriptives contribuent le plus à distinguer les communautés détectées. Cette approche permet non seulement d'évaluer l'interprétabilité des clusters, mais aussi de mesurer dans quelle mesure les comportements observés recoupent ou au contraire dépassent les catégorisations éditoriales existantes.

Enfin, nous reconnaissons que l'interprétation des clusters demeure partiellement subjective, comme dans toute tâche de clustering non supervisé. Les résultats présentés doivent donc être compris comme une exploration des comportements plausibles présents dans les données plutôt que comme une segmentation définitive des utilisateurs.

4 Étude de cas : La plateforme OpenEdition

Pour démontrer l'applicabilité de notre *framework*, nous l'avons appliqué aux données d'utilisation de la plateforme OpenEdition, une infrastructure numérique complète dédiée à l'édition en libre accès dans le domaine des sciences humaines et sociales (SHS). OpenEdition gère trois plateformes principales : *OpenEdition Journals* (revues), *OpenEdition Books* (livres) et *Hypothèses* (carnets de recherche).

Notre analyse porte sur les fichiers logs de ces trois plateformes sur une période d'un mois, totalisant 35 760 675 requêtes. Chaque requête est liée à une ressource spécifique (article, chapitre de livre ou billet de blog). Nous enrichissons ces données avec des métadonnées décrivant les ressources consultées : type de ressource, discipline (selon le système de classification interne d'OpenEdition), date de publication et pays de l'éditeur. De plus, nous utilisons une base de données de géolocalisation externe pour déterminer l'origine géographique des requêtes à partir des adresses IP.

Dans cette étude, la valeur $\alpha = 5$ a été retenue comme choix approprié, sur la base de la stabilisation observée du nombre de sessions lors des essais successifs (voir Figure 1).

Afin d'identifier la structure de partitionnement la plus robuste, nous avons effectué une recherche par grille (grid search) sur trois hyperparamètres clés : le seuil de filtrage $\beta \in [0,001; 0,005]$, le nombre de composantes NMF $k \in \{50, 80, 100\}$ et la taille minimale des clusters $ms \in [10, 25, 50, 100]$. Le tableau 1 synthétise les meilleures configurations obtenues pour chaque dimension de l'espace latent.

TABLE 1 – Résultats optimaux de la recherche par grille pour différentes dimensions NMF pour la semaine 1.

β	**k**	**ms**	**NbC**	**NR**	**BCR**	**Silh**	**DBI**
0,0035	50	25	104	0,2392	0,0842	**0,4548**	1,1142
0,0030	80	25	111	0,2781	0,0821	0,4176	1,1835
0,0020	100	10	223	0,3362	0,0581	0,3887	**1,1017**

k : composantes NMF, **ms** : *min cluster size*, **NbC** : nombre de clusters, **NR** : ratio de bruit, **BCR** : ratio du plus grand cluster, **Silh** : score de Silhouette, **DBI** : indice de Davies–Bouldin.

Pour évaluer la cohérence entre les clusters thématiques et les métadonnées documentaires, nous avons mis en place un processus d'alignement en plusieurs étapes. Chaque variable de métadonnée est d'abord transformée en un vecteur binaire par encodage *one-hot* : le type de ressource, le pays de publication, la décennie de publication, la langue, la discipline (issue de la nomenclature OE) et

le pays de session sont ainsi chacun représentés par un vecteur indicateur sur l'ensemble de leurs modalités observées. Ces vecteurs sont ensuite concaténés pour former une représentation vectorielle unique $\mathbf{x}_i \in \{0,1\}^p$ pour chaque document i, où p désigne la dimension totale de l'espace de métadonnées. Un modèle de forêt aléatoire (*Random Forest*) de 200 arbres a alors été entraîné sur ces représentations afin de prédire l'appartenance de chaque document à l'un des clusters identifiés.

TABLE 2 – Performance de classification pour différents clusters représentatifs (pour la semaine 1).

Cluster	Precision	Recall	F1-score	Support
61	0.98	0.98	0.98	334
5	0.99	0.99	0.99	174
29	0.48	0.73	0.58	15
69	0.03	0.06	0.04	34

Les résultats présentés dans le tableau 2 illustrent l'hétérogénéité des performances du modèle selon les clusters identifiés. Les clusters 61 et 5 présentent des scores de précision, de rappel et de F1 très élevés (respectivement 0,98 et 0,99), indiquant que ces groupes sont presque parfaitement prédits à partir des métadonnées disponibles. Leur taille relativement importante suggère l'existence de communautés bien structurées dont le comportement de consultation est fortement corrélé aux caractéristiques déclaratives (par exemple une discipline ou une origine géographique). À l'inverse, le cluster 29 obtient un score F1 plus modéré (0,58), révélant une structure comportementale détectée par le clustering mais moins facilement expliquée par les métadonnées. Enfin, le cluster 69 présente des performances très faibles, ce qui suggère l'existence d'un groupe principalement défini par des motifs de navigation plutôt que par des variables descriptives classiques. Ces résultats confirment que certaines communautés d'usage sont bien capturées par les métadonnées plus que d'autres.

Pour analyser les groupes obtenus, chaque paire (cluster k, feature j), nous calculons l'importance moyenne à partir des valeurs SHAP issues d'un `TreeExplainer`, formant une matrice dans $R^{K\times p}$ normalisée sur $[0,1]$ par colonne.

La figure 2 montre l'importance moyenne des principales métadonnées pour les clusters bien prédits (F1-score $\geq 0{,}8$). Plusieurs blocs d'intensité apparaissent, indiquant que certains clusters sont fortement associés à des variables spécifiques, notamment des métadonnées géographiques, disciplinaires ou temporelles. Ces motifs visuels suggèrent l'existence de communautés d'usage cohérentes, caractérisées par des profils de consultation similaires. À l'inverse, quelques clusters présentent une importance plus diffuse des variables, ce qui indique des comportements moins directement expliqués par les métadonnées. Dans l'ensemble, cette matrice confirme que de nombreux clusters correspondent à des regroupements interprétables à partir des caractéristiques descriptives des articles.

Plus précisément, plusieurs communautés semblent structurées autour de disciplines académiques identifiables, telles que l'histoire, la psychologie, l'éducation, l'économie ou encore l'ethnologie, traduisant des comportements de navigation cohérents avec des centres d'intérêt thématiques spécifiques. D'autres clusters apparaissent davantage liés à des dimensions linguistiques ou géographiques, avec des regroupements associés à certaines langues ou à certains pays. Enfin, l'hétérogénéité des motifs observés suggère que certaines communautés reposent sur des comportements plus transversaux ou émergents, moins directement capturés par les métadonnées disponibles, ce qui met en évidence les limites explicatives des métadonnées pour caractériser l'ensemble des usages observés.

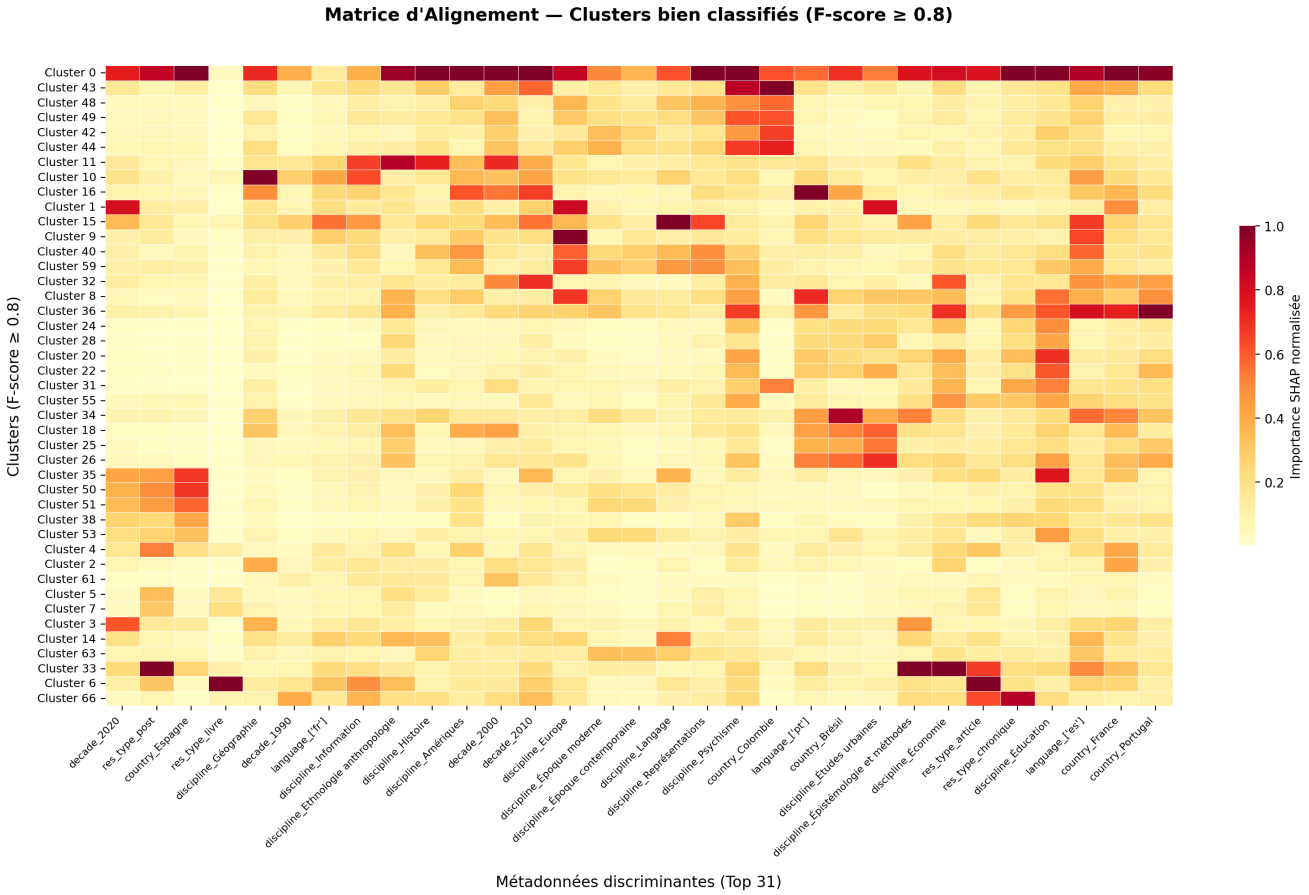

FIGURE 2 – Matrice d'alignement pour la semaine 1

5 Discussion

Ce travail propose un cadre méthodologique permettant d'exploration de communautés comportementales latentes et d'évaluation de leur alignement avec des dimensions descriptives externes. En analysant les patterns de consultation au niveau des sessions et en les reliant aux métadonnées des contenus, il devient possible de faire émerger des groupes d'intérêt partageant des caractéristiques similaires. Les items consultés doivent ainsi être compris comme des *proxys* des intérêts des utilisateurs. Dans le contexte d'une bibliothèque numérique académique comme OpenEdition, cette approche permet d'esquisser une cartographie de certaines communautés de lecture et d'observer comment leurs pratiques se structurent autour de dimensions telles que les disciplines, les aires géographiques ou les temporalités éditoriales. Certains clusters apparaissent relativement persistants et bien alignés avec les métadonnées, tandis que d'autres semblent refléter des formes d'usage plus transversales, moins directement explicables par les catégories éditoriales classiques.

Cette approche présente plusieurs avantages : elle permet d'explorer les structures d'usage d'une plateforme à grande échelle et d'identifier des communautés d'intérêt à partir des interactions réelles des utilisateurs, ce qui peut contribuer à améliorer la compréhension des usages et potentiellement informer la conception de systèmes de recommandation. Toutefois, plusieurs limites doivent être soulignées. Les métadonnées disponibles ne capturent qu'une partie des caractéristiques susceptibles d'expliquer les comportements observés, et les relations mises en évidence relèvent de corrélations plutôt que de causalité. De plus, les résultats peuvent dépendre du domaine étudié et de la structure spécifique des interactions présentes dans les données, tandis que la représentation des comportements sous forme d'embeddings implique nécessairement une simplification des pratiques de navigation. Malgré ces limites, cette approche ouvre des perspectives intéressantes pour l'étude des communautés d'utilisateurs dans les environnements web complexes, ainsi que pour l'analyse des dynamiques de lecture et des pratiques scientifiques dans les bibliothèques numériques.

Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du Programme d'Investissements d'Avenir portant la référence ANR-21-ESRE-0045.

Références

AGOSTI M., CRIVELLARI F. & DI NUNZIO G. M. (2012). Web log analysis : a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining and Knowledge Discovery*, **24**(3), 663–696. DOI : 10.1007/s10618-011-0228-8.

ALOKLUK J. A. & AL-AMRI A. (2021). Evaluation of a Digital Library : An Experimental Study. *Journal of Service Science and Management*, **14**(01), 96–114. DOI : 10.4236/jssm.2021.141007.

BOGAARD T. (2018). On the Interplay Between Search Behavior and Collections in Digital Libraries and Archives. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - CHIIR '18*, p. 339–341, New Brunswick, NJ, USA : ACM Press. DOI : 10.1145/3176349.3176350.

FU Y., LOMAS E. & INSKIP C. (2021). Library log analysis and its implications for studying online information seeking behavior of cultural groups. *The Journal of Academic Librarianship*, **47**(5), 102421. DOI : 10.1016/j.acalib.2021.102421.

HALFAKER A., KEYES O., KLUVER D., THEBAULT-SPIEKER J., NGUYEN T., GRANDPREY-SHORES K., UDUWAGE A. & WARNCKE-WANG M. (2015). User Session Identification Based on Strong Regularities in Inter-activity Time. In *Proceedings of the 24th International Conference on World Wide Web*, p. 410–418, Florence Italy : International World Wide Web Conferences Steering Committee. DOI : 10.1145/2736277.2741117.

HIDASI B., KARATZOGLOU A., BALTRUNAS L. & TIKK D. (2016). Session-based recommendations with recurrent neural networks.

ILIOU C., KOSTOULAS T., TSIKRIKA T., KATOS V., VROCHIDIS S. & KOMPATSIARIS Y. (2019). Towards a framework for detecting advanced Web bots. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, p. 1–10, Canterbury CA United Kingdom : ACM. DOI : 10.1145/3339252.3339267.

JAGAT R. R., SISODIA D. S. & SINGH P. (2024). Exploiting web content semantic features to detect web robots from weblogs. *Journal of Network and Computer Applications*, **230**, 103975. DOI : https ://doi.org/10.1016/j.jnca.2024.103975.

JAMALI H. R., NICHOLAS D. & HUNTINGTON P. (2005). The use and users of scholarly e-journals : a review of log analysis studies. *Aslib Proceedings*, **57**(6), 554–571. DOI : 10.1108/00012530510634271.

KANG W.-C. & MCAULEY J. (2018). Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, p. 197–206. DOI : 10.1109/ICDM.2018.00035.

KOVACEVIC A., DEVEDZIC V. & POCAJT V. (2010). Using data mining to improve digital library services. *The Electronic Library*, **28**(6), 829–843. DOI : 10.1108/02640471011093525.

LIANG S. & LENG Y. (2020). Search Topic Analysis of ACM Digital Library. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, p. 487–488, Virtual Event China : ACM. DOI : 10.1145/3383583.3398576.

NICHOLAS D., HUNTINGTON P. & JAMALI H. R. (2008). User diversity : as demonstrated by deep log analysis. *The Electronic Library*, **26**(1), 21–38. DOI : 10.1108/02640470810851716.

PAPATHEODOROU C., KAPIDAKIS S., SFAKAKIS M. & VASSILIOU A. (2003). Mining User Communities in Digital Libraries. *Information Technology and Libraries*, **22**.

TRABELSI M., SUIRE C., MORCOS J. & CHAMPAGNAT R. (2021). A New Methodology to Bring Out Typical Users Interactions in Digital Libraries. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, p. 11–20, Champaign, IL, USA : IEEE. DOI : 10.1109/JCDL52503.2021.00013.

WU S., TANG Y., ZHU Y., WANG L., XIE X. & TAN T. (2019). Session-based recommendation with graph neural networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19 : AAAI Press. DOI : 10.1609/aaai.v33i01.3301346.

ZERHOUDI S., GRANITZER M., SEIFERT C. & SCHLÖTTERER J. (2022). Simulating User Interaction and Search Behaviour in Digital Libraries. In G. M. D. NUNZIO, B. PORTELLI, D. REDAVID & G. SILVELLO, Éds., *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022 (hybrid event)*, volume 3160 de *CEUR Workshop Proceedings* : CEUR-WS.org.