Fouille Interactive de Motifs à Haute Utilité

Arnold Hien**, Samir Loudni**, Maxime Garfagni*,
Albrecht Zimmermann***, Nicolas Beldiceanu**

*IMT Atlantique, Brest, France
maxime.garfagni@imt-atlantique.net
** TASC – DAPI, IMT-Atlantique, LS2N – CNRS, Nantes, France
{arnold.hien, samir.loudni, nicolas.beldiceanu}@imt-atlantique.fr
***CNRS - UMR GREYC, Normandie Univ., UNICAEN, Caen, France
albrecht.zimmermann@unicaen.fr

Résumé. Ces dernières années, l'extraction de motifs à haute utilité (ou HUI, High Utility Itemset) a fait l'objet de nombreux travaux. La fouille des HUIs permet en effet la découverte de motifs dont les items présentent des pondérations traduisant leur importance dans les transactions qui les contiennent. Les HUIs sont alors des motifs présentant une utilité ou une importance certaine pour l'utilisateur. Dans les différentes approches de l'état de l'art, leur extraction nécessite de fournir en plus du jeu de données, un ensemble d'utilités associées aux items et aux transactions. Si l'existence du jeux de données est une évidence, celle des utilités peut éventuellement être problématique dans certaines situations. Dans cet article, nous proposons une méthode interactive d'extraction de motifs à haute utilité. Notre approche permet d'apprendre les utilités des items et d'améliorer itérativement l'ensemble des HUIs extraits.

1 Introduction

Ces dernières années, plusieurs travaux se sont intéressés à la fouille de HUIs (Yao et al., 2004) qui consiste à trouver des motifs dits à haute utilité dans les jeux de données. Cette utilité peut être mesurés en terme de risque, profit, poids, quantité ou autre en fonction des préférences de l'utilisateur. La fouille des HUIs peut alors être vue comme une généralisation de la fouille de motifs fréquents dans lequel les items de chaque transaction sont associés à un poids (utilité externe) représentant leur importance, ainsi qu'à une utilité interne représentant une caractéristique associé comme la quantité (Yao et al., 2004; Yao et Hamilton, 2006). Cette approche de fouille peut ainsi être utilisée dans les applications quotidiennes comme l'analyse des flux de clics (Chu et al., 2008), des achats (Li et al., 2008) et comportements des utilisateurs (Shie et al., 2013) ou dans la génétique (Zihayat et al., 2017).

Différentes méthodes plus au moins efficaces ont été proposées pour l'extraction de HUIs (Liu et Qu, 2012; Fournier-Viger et al., 2015; Tseng et al., 2016; Qu et al., 2019; Duong et al., 2022; Diop, 2022), mais elles nécessitent toutes de définir au préalable les utilités des items au moment de la fouille. Cependant, déterminer l'utilité des items n'est pas trivial, car souvent définies de manière arbitraire, rendant leur exploitation difficile dans un contexte applicatif.

Pour résoudre ce problème, nous proposons de bénéficier du cadre de la fouille interactive de motifs((Dzyuba et al., 2014; Dzyuba et van Leeuwen, 2017; Hien et al., 2023b)) pour apprendre les utilités attribuées aux items selon les retours de l'utilisateur. Dans cet article, nous proposons une approche interactive de fouille d'HUIs permettant à la fois d'apprendre les préférences de l'utilisateur et d'extraire les HUIs selon ses préférences. Cette approche présente l'avantage de s'affranchir de la définition initiale de seuil d'utilité par l'utilisateur et permet d'intégrer ses connaissances dans les valeurs des utilités. A notre connaissance, il s'agit de la toute première approche interactive pour l'extraction d'HUIs.

2 Préliminaires

Dans cette section nous présentons quelques préliminaires sur la fouille d'HUIs et sur l'apprentissage des préférences de l'utilisateur dans un contexte interactif.

Fouille de HUIs (Motifs à Haute Utilité). Soit $\mathcal{I} = \{1, \ldots, n\}$ un ensemble de n items et $\mathcal{T} = \{1, ..., m\}$ un ensemble de m indices de transactions. Un jeu de données transactionnelles $\mathcal{D}$ est un ensemble de transactions, où chaque transaction t est un sous-ensemble de $\mathcal{I}$ et possède un identifiant unique $r \in \mathcal{T}$. On définit un motif X comme un sous-ensemble non vide de $\mathcal{I}$ et sa couverture $\mathcal{V}_{\mathcal{D}}(X)$ est égale à l'ensemble des transactions qui le supportent, *i.e.,* $\mathcal{V}_{\mathcal{D}}(X) = \{t \in \mathcal{D} \mid X \subseteq t\}$. Dans cet article, chaque item i d'une transaction t est associé à un poids $w(i) \in \mathbb{R}^+$, appelé *utilité externe* et représentant son importance relative pour l'utilisateur. Par ailleurs, chaque item i d'une transaction t est associé à un entier positif $q(i,t)$, appelé *utilité interne*, représentant la quantité de i dans t. L'utilité $u(i,t)$ d'un item i dans une transaction t est définie par $u(i,t) = w(i) \times q(i,t)$. L'utilité d'un motif X dans une transaction t, notée $u(X,t)$, est définie par $u(X,t) = \sum_{i \in X} u(i,t)$ si $X \subseteq t$, et $u(X,t) = 0$ sinon. L'utilité de X dans le jeu de données, notée $u(X)$, est définie par : $u(X) = \sum_{t \in \mathcal{V}_{\mathcal{D}}(X)} u(X,t)$. En supposons que $\forall i \in \mathcal{I} \wedge \forall t \in \mathcal{D} : q(i,t) = 1$, nous avons alors $u(X) = |\mathcal{V}_{\mathcal{D}}(X)| \cdot \sum_{i \in X} w(i)$. On dit que X est un motif à haute utilité si son utilité $u(X)$ est supérieure ou égale à un seuil d'utilité minimale θ_{hui} (*i.e.,* $u(X) \geq \theta_{hui}$).

La tâche d'extraction de HUIs consiste alors à trouver tous les motifs X tels que $u(X) \geq \theta_{hui}$ (Yao et al., 2004). Les méthodes de l'état de l'art font cependant face à deux difficultés majeures : la non-monotonie de la mesure d'utilité (Yao et al., 2004) qui rend son utilisation difficile pour réduire l'espace de recherche, et le problème d'explosion des motifs qui rend difficile leur analyse. Différentes approches TOP-k ont alors été proposées (Chan et al., 2003; Tseng et al., 2016) pour s'affranchir du seuil θ_{hui} et pour contrôler la taille de sortie. TKO et TKU (Tseng et al., 2016) font partie des approches les plus connues pour l'extraction des TOP-k HUIs et se différencient par la structure des données utilisée. Cependant, malgré leur capacité à extraire l'ensemble des TOP-k HUIs, elles demeurent peu efficaces notamment avec des jeux de données de grandes tailles. Pour apporter plus d'efficacité, des heuristiques comme TKU-CE (Song et al., 2020) et des méthodes d'échantillonnage, comme HAISAMPLER (Diop, 2022) ont été proposées. Dans TKU-CE, les TOP-k HUIs sont extraits de manière itérative avec un vecteur de probabilité initialisé de manière aléatoire et mis à jour à chaque itération pour tirer de nouveaux échantillons, tandis que HAISAMPLER permet de tirer les HUIs proportionnellement à leur utilité moyenne en respectant une contrainte de taille.Toutes ces méthodes nécessitant une définition préalable des utilités externes des items, nous proposons d'exploiter le cadre de la fouille interactive de motifs pour apprendre ces utilités.

Algorithme 1 : LUTOM-HUI (**L**earn **U**tilities for **To**p-k **HUI**s **M**ining)

Input : jeu de données $\mathcal{D}$
Parameters : taille de requête k, nombre d'itérations T, descripteurs des motifs $\mathcal{F}$,
ORACLE d'extraction des HUIs, par. de rétention d'une requête ℓ
DISCRIMINANT : paramètre bool. pour indiquer la présence ou pas de discriminants
Output : φ : fonction d'apprentissage
Begin
  $\mathcal{U} \leftarrow \emptyset$, $w^0_{\mathcal{F}} \leftarrow$ utilités initiales, $u^0 \leftarrow$ fonction d'utilité initiale ▷ Initialisations
  pour $t = 1, 2 \dots T$
    $\mathcal{X}^t \leftarrow \text{TOP}(\mathcal{X}^{t-1}, \ell) \cup \text{ORACLE}(\mathcal{D}, (k-\ell), u^{t-1})$ ▷ Extraction des TOP-k
    $\hat{\mathcal{R}}^t \leftarrow \text{RANGEMENT}(\mathcal{X}^t)$ ▷ Interaction avec l'utilisateur
    $\mathcal{U} \leftarrow \mathcal{U} \cup \hat{\mathcal{R}}^t$
    si DISCRIMINANT **alors**
      $disc \leftarrow \text{EXTRAIRE-DISCRIMINANT}(\mathcal{X}^t, \hat{\mathcal{R}}^t)$ ▷ Extraction du discriminant
      $\langle w^t_{\mathcal{F}}, w^t_{\mathcal{F}_{disc}} \rangle \leftarrow \text{APPRENTISSAGE}(\mathcal{U}, \mathcal{F} \cup \mathcal{F}_{disc})$ ▷ Apprentissage des utilités
      $w^t_{\mathcal{F}} \leftarrow \text{AGREG}(w^t_{\mathcal{F}}, w^t_{\mathcal{F}_{disc}})$
    sinon
      $w^t_{\mathcal{F}} \leftarrow \text{APPRENTISSAGE}(\mathcal{U}, \mathcal{F})$ ▷ Apprentissage des utilités
    $u^t \leftarrow f^o(w^t_{\mathcal{F}})$ ▷ mis à jour de u^t à partir de $w^t_{\mathcal{F}}$
  return φ;

Fouille interactive de motifs. Étant donné un jeu de données $\mathcal{D}$ et un langage de motifs $\mathcal{L}_{\mathcal{I}}$, nous considérons une fonction inconnue $u^* : \mathcal{L}_{\mathcal{I}} \rightarrow \mathbb{R}$ représentant les utilités des motifs. Cette fonction donne une estimation numérique de l'intérêt des motifs et permet d'approximer les préférences de l'utilisateur encodées par $\succ_{u^*} : \forall X, Y \in \mathcal{L}_{\mathcal{I}}$, $u^*(X) > u^*(Y)$ ssi $X \succ_{u^*} Y$. Ainsi, les préférences de l'utilisateur peuvent prendre la forme d'un ordre total $\mathcal{R}^*$ sur les motifs $X_i \in \mathcal{L}_{\mathcal{I}} : X_{\pi(1)} \succ_{u^*} \cdots \succ_{u^*} X_{\pi(n)}$ avec $\pi(i)$ l'indice du motif de rang i. Le calcul des utilités des motifs s'effectue en exploitant une combinaison de poids $w_{\mathcal{F}}$ associés à des descripteurs $\mathcal{F}$ que nous utilisons pour représenter les motifs. Cette représentation des motifs se présente sous forme vectorielle où chaque F_i désigne une caractéristique à représenter. Les poids $w_{\mathcal{F}}$ représentent alors les utilités externes de chaque descripteur et indiquent l'importance des caractéristiques associées. Dans cet article, nous proposons d'apprendre les poids $w_{\mathcal{F}}$ permettant d'obtenir une fonction d'utilité u correspondant aux préférences de l'utilisateur. Étant donné la difficulté pour un utilisateur d'attribuer des utilités numériques précises aux différents motifs, nous proposons un processus d'apprentissage suivant le framework MINE-INTERACT-LEARN (Dzyuba et van Leeuwen, 2017; Hien et al., 2023b) :
(1) **MINE** : un ensemble (suffisamment réduit) de HUIs est extrait en utilisant la fonction d'utilité apprise u ainsi que les utilités apprises des descripteurs $w_{\mathcal{F}}$;
(2) **INTERACT** : l'utilisateur est invité à ranger les HUIs extraits selon ses préférences ;
(3) **LEARN** : Le retour de l'utilisateur est exploité pour mettre à jour les poids $w_{\mathcal{F}}$;
(4) **MINE** (encore) : un nouvel ensemble de HUIs est extrait en exploitant la fonction mise à jour u. Le processus se poursuit entre les étapes 2 et 4 jusqu'à l'atteinte d'un critère d'arrêt.

3 Apprentissage interactif des utilités des items

Dans cette section, nous présentons LUTOM-HUI, une instanciation du framework MINE-INTERACT-LEARN pour la fouille interactive des TOP-k HUIs, dont le pseudo-code est donné

par l'algorithme 1. LUTOM-HUI maintient une approximation u^t de la fonction d'utilité u^*, où t est l'indice d'itération. À chaque itération t, l'algorithme sélectionne une requête $\mathcal{X}^t$ à poser à l'utilisateur (ligne 9), et recueille ses préférences $\widehat{\mathcal{R}}^t$ (ligne 10) qui sont ensuite utilisées pour calculer une nouvelle approximation u^t de la fonction u^* (ligne 18). Le processus se poursuit pendant T itérations au bout desquelles l'algorithme retourne l'approximation finale u^T. Cette section est consacrée à la description des différentes composantes de LUTOM-HUI. Les notations utilisées sont définies dans la Table 1b.

A) Descripteurs des motifs. L'ensemble des descripteurs des motifs forment un vecteur $\mathcal{F} = \langle F_1, \ldots, F_n \rangle$ (avec n le nombre de descripteurs), où chaque F_i représente une caractéristique des motifs. Dans LUTOM-HUI, nous donnons la possibilité d'utiliser deux groupes de descripteurs dits *statiques* : des descripteurs pour les items, notés $\mathcal{F}_\mathcal{I}$, et pour les transactions, notés $\mathcal{F}_\mathcal{T}$. Chaque motif X est alors représenté par un vecteur $X_\mathcal{F} = \langle X_{F_1}, \ldots, X_{F_n} \rangle$, où X_{F_i} représente la valeur du descripteur F_i. En utilisant $\mathcal{F}_\mathcal{I}$, les motifs sont représentés par un vecteur de valeurs binaires $X_\mathcal{I}$ avec $X_{\mathcal{I}_i} = 1$ si $i \in X$ et 0 sinon. En combinant $\mathcal{F}_\mathcal{I}$ et $\mathcal{F}_\mathcal{T}$, nous obtenons également un vecteur de valeurs binaires $X_\mathcal{F} = \langle X_\mathcal{I}, X_\mathcal{T} \rangle = \langle X_{\mathcal{I}_1} \ldots X_{\mathcal{I}_{|\mathcal{I}|}}, X_{\mathcal{T}_1} \ldots X_{\mathcal{T}_{|\mathcal{T}|}} \rangle$, avec $X_{\mathcal{T}_t} = 1$ si $t \in \mathcal{V}_\mathcal{D}(X)$ et 0 sinon.

Afin d'apporter plus d'expressivité à la représentation des motifs, nous proposons également l'utilisation des descripteurs discriminants (Hien et al., 2023b) qui mettent en exergue les relations entre les items. Ces descripteurs exploitent la notion de *motifs discriminants* qui sont des motifs corrélés au rangement $\widehat{\mathcal{R}}$. Ainsi, étant donné $disc$ le motif discriminant extrait à l'itération t (ligne 13) et $\mathcal{F}_{disc}$ le descripteur discriminant associé, nous obtenons la description X_{disc} telle que $X_{disc} = 1$ si $disc \subseteq X$ et 0 sinon. Nous introduisons alors de nouveaux *descripteurs temporaires* $\mathcal{F}^*$ afin de bénéficier de la sémantique apportée par les discriminants : $\mathcal{F}^* = \langle \mathcal{F}_\mathcal{I}, \mathcal{F}_\mathcal{T}, \mathcal{F}_{disc} \rangle$ et $X_{\mathcal{F}^*} = \langle X_{\mathcal{I}_1} \ldots X_{\mathcal{I}_{|\mathcal{I}|}}, X_{\mathcal{T}_1} \ldots X_{\mathcal{T}_{|\mathcal{T}|}}, X_{disc} \rangle$.

B) Fonction d'utilité. Étant donné $w_\mathcal{F}$ le vecteur de poids associés aux descripteurs $\mathcal{F}$, la fonction d'utilité u^t se présente sous forme d'une combinaison pondérée des descripteurs F_i de $\mathcal{F}$, où chaque poids w_{F_i} représente l'utilité externe et la contribution de F_i à l'utilité du motif. Notre fonction d'utilité s'exprime alors comme suit, en fonction des descripteurs utilisés : (1) lorsque $\mathcal{F} = \mathcal{F}_\mathcal{I}$ (I) : $u^t(X) = |\mathcal{V}_\mathcal{D}(X)| \cdot (w^T_{\mathcal{F}_\mathcal{I}} \cdot X_\mathcal{I})$; (2) lorsque $\mathcal{F} = \langle \mathcal{F}_\mathcal{I}, \mathcal{F}_\mathcal{T} \rangle$ (IT) : $u^t(X) = (w^T_{\mathcal{F}_\mathcal{I}} \cdot X_\mathcal{I}) \cdot (w^T_{\mathcal{F}_\mathcal{T}} \cdot X_\mathcal{T})$, avec $w_{\mathcal{F}_\mathcal{I}}$ (resp. $w_{\mathcal{F}_\mathcal{T}}$) les poids associés à $\mathcal{F}_\mathcal{I}$ (resp. $\mathcal{F}_\mathcal{T}$).

C) Apprentissage à partir des feedback. L'objectif de la composante LEARN est d'apprendre des poids $w_\mathcal{F}$ permettant d'obtenir une fonction d'utilité u^t correspondant aux préférences de l'utilisateur. Cette apprentissage exploite le rangement $\widehat{\mathcal{R}}^t$ (ligne 10) pour apprendre les poids $w^t_\mathcal{F}$ des descripteurs statiques $\mathcal{F}$ et le poids $w^t_{\mathcal{F}_{disc}}$ du descripteur discriminant $\mathcal{F}_{disc}$ (lignes 14 et 17). Ainsi, étant donné un feedback utilisateur $\mathcal{U} = \{X_1 \succ \cdots \succ X_k\}$ qui se traduit en classements par paires $\{(X_1 \succ X_2), (X_1 \succ X_k), \ldots\}$, chaque paire classée $X_i \succ_{u^t} X_j$ correspond à un exemple de classification $(X_{F_i} - X_{F_j}, +)$ d'un ensemble d'apprentissage. L'apprentissage consiste alors à résoudre un problème de classification dont l'objectif est de minimiser les discordances entre $\hat{\mathcal{R}}$ et $\mathcal{R}^*$. Pour cela, LUTOM-HUI utilise la méthode SCD (Shalev-Shwartz et Tewari, 2011). A la fin, nous obtenons de nouveaux poids $w^t_\mathcal{F}$ (ligne 17) qui serviront à mettre à jour u^t (ligne 18) afin d'extraire de nouveaux HUI à l'itération suivante.

Deux situations se présentent alors en fonction de la présence ou pas de discriminants. En présence de descripteur discriminant, l'apprentissage réalisé à la ligne 14 permet de mettre à jour les poids $w^t_\mathcal{F}$ et $w^t_{\mathcal{F}_{disc}}$. Étant donné que le discriminant est utilisé comme descripteur

TAB. 1 – Jeux de données et notations.

(a) Dataset Characteristics.

Datasets	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{I}\rvert$
Chess	3196	75
Foodmart	4 141	1559
German-credit	1 000	110
Mushroom	8 124	112
Soybean	630	50
Vote	435	48

(b) Notations.

Notation	Signification	Notation	Signification
$t \in [T]$	indice des Itérations	$\mathcal{F}_{disc}$	descripteur discriminant
$\mathcal{X}^t$	Requête utilisateur à l'itération t	$X_{\mathcal{F}}$	représentation du motif X avec $(\mathcal{F} \cup \mathcal{F}_{disc})$
$\hat{\mathcal{R}}^t$	Rangement de l'utilisateur sur $\mathcal{X}^t$	$disc$	motif discriminant extrait à partir de $\widetilde{R}^t$
$\mathcal{F}$	ensemble des descripteurs	$w_{\mathcal{F}}{}^t$	poids associés aux descripteurs statiques $\mathcal{F}$
$\mathcal{F}_{\mathcal{I}}$	descripteurs des items	$w_{\mathcal{F}_{disc}}$	poids associés au descripteur discriminant $\mathcal{F}_{disc}$
$\mathcal{F}_{\mathcal{T}}$	descripteurs des transactions	u	fonction d'utilité

temporaire, le poids $w^t_{\mathcal{F}_{disc}}$ appris est agrégé aux autres poids de deux manières différentes (Hien et al., 2023b) : (i) avec une fonction multiplicative d'agrégation linéaire (LIN) : $w^t_{F_i} = w^t_{F_i} \times (1 + \eta \cdot w^t_{F_{disc}})$ (ii) avec une fonction multiplicative d'agrégation exponentielle (EXP) : $w^t_{F_i} = w^t_{F_i} \times \exp^{\eta \cdot w^t_{F_{disc}}}$. Le paramètre $\eta \in]0, \frac{1}{2}]$ est un paramètre de régularisation pour limiter les fortes variations des poids résultant de cette deuxième mise à jour. En l'absence de discriminants, un simple apprentissage est réalisée (ligne 17).

D) Extraction des HUIs. La sélection des motifs est une étape essentielle de la fouille interactive car elle permet d'améliorer l'apprentissage et de réduire l'effort d'analyse de l'utilisateur. Pour cela, notre approche consiste à sélectionner dans la requête $\mathcal{X}$ les k HUIs les plus pertinents selon la fonction d'utilité u. Ainsi, à chaque itération t, nous exploitons les poids $w^t_{\mathcal{F}}$ ainsi que la fonction d'utilité u^t apprise à l'itération précédente pour sélectionner les TOP-k HUIs à présenter à l'utilisateur (ligne 9). Cette extraction est alors réalisée avec les oracles TKO, TKU-CE et HAISAMPLER. Notons que dans le cas où nous utilisons les descripteurs IT (voir 3-B), nous introduisons la possibilité de capturer la contribution des transactions dans l'utilité des HUIs. Enfin, pour permettre à l'utilisateur de faire le lien entre les itérations, la requête $\mathcal{X}^t$ est construite en retenant les ℓ meilleurs HUIs rangés par l'utilisateur à l'itération $(t-1)$ auxquels s'ajoutent les TOP-(k-ℓ) HUIs extraits avec l'oracle.

4 Évaluation expérimentale

a) Protocole expérimental. L'évaluation de LUTOM-HUI consiste à mesurer la précision et l'efficacité de l'apprentissage avec des oracles d'extraction de HUIs. Pour cela, nous émulons l'utilisateur avec une fonction d'utilité u^* utilisée comme fonction de rangement des HUIs. u^* exploite un ensemble d'utilités externes $w^*_{\mathcal{F}}$ générées suivant une distribution Normale, et est définie de la manière suivante, en fonction du descripteur utilisé : $u^*(X) = |\mathcal{V}_{\mathcal{D}}(X)| \cdot (X_{\mathcal{I}} \cdot w^*_{\mathcal{F}_{\mathcal{I}}})$ si $\mathcal{F} = \mathcal{F}_{\mathcal{I}}$ (I); $u^*(X) = (X_{\mathcal{I}} \cdot w^*_{\mathcal{F}_{\mathcal{I}}}) \cdot (X_{\mathcal{T}} \cdot w^*_{\mathcal{F}_{\mathcal{T}}})$ si $\mathcal{F} = \langle \mathcal{F}_{\mathcal{I}}, \mathcal{F}_{\mathcal{T}} \rangle$ (IT).

Pour mesurer les performances de LUTOM-HUI, nous utilisons la mesure du *regret* qui évalue sa capacité à sélectionner des motifs avec une grande utilité. À chaque itération t, nous calculons le rang centile $pct.rank(X_i)$ des motifs $X_i \in \mathcal{X}^t$ comme suit : $pct.rank(X_i) = (DI + \frac{DE}{2})/|\mathrm{S}|$, avec $DI = |Y \in \mathrm{S},\ u^*(Y) < u^*(X_i)|$, $DE = |Y \in \mathrm{S}, u^*(Y) = u^*(X_i)|$ et S représente les TOP-10^4 HUIs extraits avec $w^*_{\mathcal{F}}$. Le regret est alors défini comme suit : $Regret_M(\mathcal{X}^t) = 1 - M_{(1 \leq i \leq k)}(pct.rank(X_i))$. Il peut être évalué en utilisant la moyenne des rangs centiles ($M = Avg$) ou leur valeur maximum ($M = Max$). Les poids $w^*(i)$ étant générés aléatoirement, chaque combinaison de paramètre est évaluée 10 fois avec dif-

férentes distributions gaussiennes ; le regret obtenu correspond à la moyenne des regrets. Pour chaque jeu de données, le regret est calculé pour les différentes itérations. Nous avons utilisé les jeux de données de l'UCI (voir Table 1a) avec les paramétrages suivants : la taille de requête $k \in \{3, 5, 7, 10\}$; les descripteurs I et IT ; le paramètre de rétention de requête $\ell \in \{0, 1\}$; les oracles $\{$TKO, TKU-CE, HAISAMPLER$\}$; deux approches d'apprentissage : LUTOM-HUI avec discriminants (noté LUTOM-DISC), et sans discriminants (noté LUTOM-SANS-DISC). Nous notons LUTOM-DISC-E (resp. LUTOM-DISC-L) l'utilisation de l'agrégation exponentielle (resp. linéaire) dans LUTOM-DISC.

b) Résultats expérimentaux. La première série de nos expérimentation a consisté à déterminer la meilleure taille de requête k ainsi que les meilleurs paramétrages pour LUTOM-DISC : la fonction d'agrégation et la valeur de η.

b1) Analyse préliminaire. Par soucis d'espace, les graphique de cette évaluation sont donnés dans Hien et al. (2023a). Ainsi, nous relevons que les meilleurs regrets sont presque toujours obtenus pour $\eta = 0.13$ et $k = 7$ quel que soit l'oracle utilisé. Par ailleurs, nous relevons que l'approche LUTOM-DISC-L donne de meilleurs résultats que LUTOM-DISC-E et que TKO et TKU-CE permettent d'obtenir de meilleures valeurs de regret que HAISAMPLER. Pour évaluer l'impact des descripteurs des motifs, nous analysons les regrets obtenus par LUTOM-HUI avec les descripteurs d'items $\mathcal{F}_{\mathcal{I}}$ et de transactions $\mathcal{F}_{\mathcal{T}}$.

b2) Évaluation de l'impact des descripteurs de motifs. Les résultats de cette analyse sont données dans la Table 2 où nous avons agrégés les regrets sur tous les jeux de données, Nous observons alors que l'oracle TKO est celui qui donne les meilleurs résultats, quelle que soit la mesure de regret considéré et quel que soit le descripteur. Ces résultats confirme donc la tendance observée dans l'analyse préliminaire. Par ailleurs, nous observons que l'ajout du descripteur $\mathcal{F}_{\mathcal{T}}$ à $\mathcal{F}_{\mathcal{I}}$ ne permet pas d'améliorer fortement les résultats de LUTOM-HUI, ce qui pourrait indiquer une difficulté à apprendre correctement le poids des transactions. Nous remarquons également que les résultats obtenus par LUTOM-SANS-DISC sont meilleurs que ceux de LUTOM-DISC. En effet, les valeurs de regrets obtenus sont presque toujours inférieures de moitié à ceux de LUTOM-DISC. Cela indique que l'ajout temporaire du descripteur $\mathcal{F}_{disc}$ n'a pas permis d'améliorer l'apprentissage et l'a probablement perturbé.

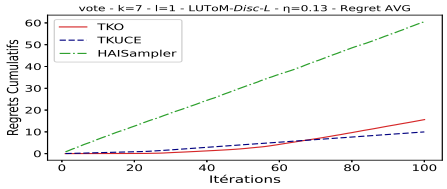
b3) Évolution du temps d'exécution et du regret par itération. La Figure 1 présente une vue détaillée des regrets cumulatifs et non cumulatifs de LUTOM-DISC sur le jeu de données VOTE (les autres résultats sont donnés dans Hien et al. (2023a)). Ces graphiques montrent l'évolution du regret (cumulatif et non cumulatif) sur 100 itérations. Ces graphiques confirment une fois de plus la capacité de l'oracle TKO à pouvoir identifier des HUIS plus intéressants, d'où un regret cumulatif ou non cumulatif plus réduit que celui des autres oracles. La Figure 1c montre l'évolution des temps de calcul LUTOM-HUI pour les trois oracles évalués. Nous constatons alors que malgré l'utilisation d'approches TOP-k comme oracle, LUTOM-DISC réussit à retourner dans des délais plus ou moins raisonnables des HUIS. Cette observation est surtout remarquable pour l'oracle TKU-CE qui effectue une recherche heuristique. Par contre, pour l'oracle TKO qui fait une recherche exhaustive, les temps de réponses sont plus longs.

5 Conclusion

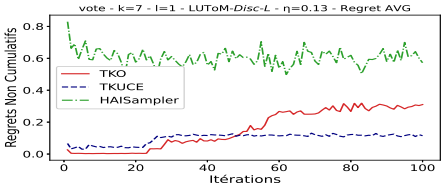
Dans cet article, nous avons introduit l'une des premières méthodes de fouille interactive de HUIS. Cette méthode exploite les descripteurs des motifs pour apprendre les utilités des items

TAB. 2 – Évaluation des descripteurs de LUTOM-HUI pour : (1)TKO, (2)TKU-CE et (3)HAISAMPLER. Les résultats sont agrégés sur tous les jeux de données pour $k = 7$ et $\ell = 1$.

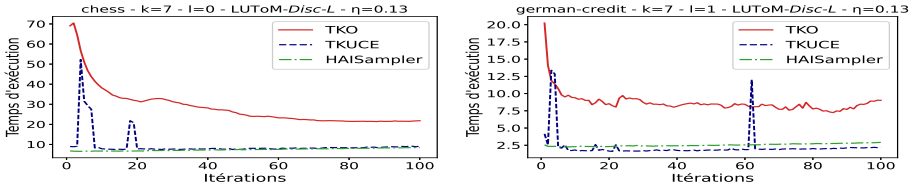
	LUTOM-DISC						LUTOM-SANS-DISC					
	$Regret_{Max}$			$Regret_{Avg}$			$Regret_{Max}$			$Regret_{Avg}$		
Descripteurs	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
I	**1.001**	1.014	2.067	2.083	**1.990**	4.064	**0.013**	0.507	4.687	**0.019**	0.353	4.067
IT	**1.000**	1.009	2.078	**1.591**	1.807	4.222	**0.116**	0.676	4.498	**0.036**	0.365	4.276



(a) VOTE : $Regret_{Avg}$ Cumulatif.



(b) VOTE : $Regret_{Avg}$ Non-Cumulatif.



(c) Temps d'exécution de CHESS et GERMAN-CREDIT

FIG. 1 – Vue détaillée des résultats de LUTOM-DISC-L avec $k = 7$, $\ell = 1$ et $\eta = 0.13$.

de manière dynamique. Les utilités apprises sont alors exploitées pour extraire des HUIs de plus en plus intéressants. Les expérimentations que nous avons menés montrent des performances encourageantes qui peuvent être améliorées notamment par l'utilisation d'un oracle plus efficace ou une fonction d'apprentissage plus précise. Ces différentes pistes d'amélioration permettront sans doute d'améliorer considérablement les résultats de LUTOM-HUI.

Références

Chan, R., Q. Yang, et Y. Shen (2003). Mining high utility itemsets. In *ICDM 2003*, pp. 19–26.

Chu, C., V. S. Tseng, et T. Liang (2008). An efficient algorithm for mining temporal high utility itemsets from data streams. *J. Syst. Softw. 81*(7), 1105–1117.

Diop, L. (2022). High average-utility itemset sampling under length constraints. In *PAKDD 2022*, pp. 134–148.

Duong, H. V., N. T. H. Le, T. Tran, T. C. Truong, B. Le, et P. Fournier-Viger (2022). Efficient algorithms for mining closed and maximal high utility itemsets. *Knowl. Based Syst. 257*.

Dzyuba, V. et M. van Leeuwen (2017). Learning what matters - sampling interesting patterns. In *PAKDD 2017, Proceedings, Part I*, pp. 534–546.

Dzyuba, V., M. van Leeuwen, S. Nijssen, et L. D. Raedt (2014). Interactive learning of pattern rankings. *International Journal on Artificial Intelligence Tools 23*(6).

Fournier-Viger, P., J. C. Lin, T. Gueniche, et P. Barhate (2015). Efficient incremental high utility itemset mining. In *ASE BD & SI 2015*, pp. 53 :1–53 :6.

Hien, A., S. Loudni, N. Aribi, A. Ouali, et A. Zimmermann (2023a). Code and supplementary material. https://gitlab.com/phdhien/dispale.

Hien, A., S. Loudni, N. Aribi, A. Ouali, et A. Zimmermann (2023b). Interactive pattern mining using discriminant sub-patterns as dynamic features. In *PAKDD 2023, Part I*, pp. 252–263.

Li, Y., J. Yeh, et C. Chang (2008). Isolated items discarding strategy for discovering high utility itemsets. *Data Knowl. Eng. 64*(1), 198–217.

Liu, M. et J. Qu (2012). Mining high utility itemsets without candidate generation. In *CIKM 2012*, pp. 55–64.

Qu, J.-F., M. Liu, et P. Fournier Viger (2019). *Efficient Algorithms for High Utility Itemset Mining Without Candidate Generation*, pp. 131–160.

Shalev-Shwartz, S. et A. Tewari (2011). Stochastic methods for l_1-regularized loss minimization. *J. Mach. Learn. Res. 12*, 1865–1892.

Shie, B., P. S. Yu, et V. S. Tseng (2013). Mining interesting user behavior patterns in mobile commerce environments. *Appl. Intell. 38*(3), 418–435.

Song, W., L. Liu, et C. Huang (2020). TKU-CE : cross-entropy method for mining top-k high utility itemsets. In *IEA/AIE 2020*, pp. 846–857.

Tseng, V. S., C. Wu, P. Fournier-Viger, et P. S. Yu (2016). Efficient algorithms for mining top-k high utility itemsets. *IEEE Trans. Knowl. Data Eng. 28*(1), 54–67.

Yao, H., H. Hamilton, et C. Butz (2004). A foundational approach to mining itemset utilities from databases. In *SIAM DM 2004*, pp. 482–486.

Yao, H. et H. J. Hamilton (2006). Mining itemset utilities from transaction databases. *Data Knowl. Eng. 59*(3), 603–626.

Zihayat, M., H. Davoudi, et A. An (2017). Mining significant high utility gene regulation sequential patterns. *BMC Syst. Biol. 11*(6), 109 :1–109 :14.

Summary

In recent years, HUI (High Utility Itemset) mining has become a substantial area of research in pattern mining. HUI Mining is a family of methods used to mine patterns by weighting items according to their importance in the transactions in which they occur. HUIs are, therefore, patterns that hold clear importance for the user. The most current approaches require the user to provide a set of utilities linked to both items and transactions in addition to the dataset. Providing this dataset is straightforward. However, obtaining the utilities can be challenging in certain scenarios. In this paper, we introduce a utility-free method for interactively mining high-utility itemsets. Our approach consists of learning the utilities from the user feedback and using the learned utilities to gradually enhance the set of HUIs.