Extraction de comportements inattendus dans le cadre du "Web Usage Mining"

Haoyuan Li, Anne Laurent, Pascal Poncelet

Extraction de comportements inattendus dans le cadre du Web Usage Mining

Dong (Haoyuan) Li*, Anne Laurent**, Pascal Poncelet**

*LGI2P, École des Mines d'Alès, Parc scientifique Georges Besse, 30035 Nîmes cedex 1
Haoyuan.Li@ema.fr
**LIRMM, Université Montpellier 2, 161 rue Ada, 34392 Montpellier cedex 5
laurent@lirmm.fr, poncelet@lirmm.fr

Résumé. Au cours de ces dernières années, la fouille de données d'usage du Web s'est de plus en plus concentrée sur l'extraction des comportements des utilisateurs à partir de fichiers logs. Bien que l'extraction de motifs séquentiels permette de trouver des comportements fréquents, les décideurs sont de plus en plus intéressés par des comportements inattendus qui contredisent les croyances sur des connaissances existantes. Dans cet article, nous présentons une nouvelle approche, WebUser, pour découvrir des comportements inattendus par rapport aux croyances du domaine. Les expérimentations menées, avec des bases de croyances générées à partir des comportements connus, montrent que notre approche permet d'extraire des comportements inattendus qui peuvent être utilisés pour, par exemple, améliorer la structure des sites Web ou repérer des usages particuliers.

1 Introduction

L'utilisation de techniques de fouille sur des données d'usage du Web pour extraire des connaissances sur les comportements des utilisateurs, i.e. le Web Usage Mining (WUM), est un domaine de recherche particulièrement étudié ces dernières années. Ainsi, de nombreuses approches ont été proposées pour analyser les logs et offrir de nouvelles connaissances utiles comme, quelles sont les ressources souvent visitées ? quel est le parcours des utilisateurs sur un site ? Ces connaissances peuvent être utilisées pour, par exemple, restructurer un site, personnaliser un site en fonction du comportement des utilisateurs ou encore précharger des pages comme les approches introduites par Büchner et Mulvenna (1998), Spiliopoulou et al. (1999), Eirinaki et Vazirgiannis (2003), Srivastava et al. (2000) et Mobasher (2007).

Parmi les techniques de fouille de données utilisées, l'extraction de *motifs séquentiels* présentée par Agrawal et Srikant (1995) est souvent utilisée car particulièrement bien adaptée à ce contexte selon les approches proposées par Mobasher et al. (2002), Masseglia et al. (2003), Huang et al. (2006), Missaoui et al. (2007), et Masseglia et al. (2007). Par exemple, via les motifs séquentiels, il est possible d'extraire des comportements du type : "sur un site Web de forum, 40% des utilisateurs visitent la page `TopicList`, puis la page `Search`, puis la page `Login`, et enfin la page `PostTopic`", ou bien dans un site d'e-commerce "80% de clients s'intéressent aux pages sur les sacoches d'ordinateur portable après avoir ajouté un PC portable

à leur panier". Ce type de connaissances permet de refléter le comportement usuel des utilisateurs sur un site Web. Cependant, en ne reflétant que le comportement fréquent il laisse un problème en suspens : quid des comportements inattendus ? En effet, les décideurs sont de plus en plus intéressés par les comportements particuliers correspondant à des niches potentielles ou à des nouveaux usages des sites. Dans ce cas, à partir des connaissances générales que le décideur peut avoir sur son site, l'objectif est de rechercher les comportements qui violent cette connaissance.

De manière à illustrer cette problématique, considérons, par exemple, un site Web de News en ligne, où les nouvelles les plus récentes apparaissent sur la page d'accueil `index.html` par catégories sous la forme d'index statiques spécifiant la catégorie (e.g. `cat1.html`, `cat2.html`, etc.). Supposons également que celles-ci puissent être obtenues via une page script dynamique `list.php` dans laquelle l'identification de la catégorie est passé en paramètre (e.g. `list.php?cat=1&page=3`) et qu'une autre page dynamique `read.php` serve à afficher les détails de la nouvelle spécifiée en paramètre (e.g. `read.php?news=080114021`). Supposons à présent que, via l'historique des accès sur le site, les comportements suivants soient trouvés : (1) 60% des utilisateurs visitent `index.html`, ensuite `read.php`, puis `cat1.html`, ensuite `read.php`, puis `cat5.html`, ensuite `read.php`, et enfin d'autres catégories via `read.php` ; (2) 10% d'utilisateurs visitent `index.html`, puis `cat5.html` et jamais la page `cat1.html`, et enfin la page `read.php` ; (3) 8% d'utilisateurs visitent `read.php` une seule fois ; (4) 0.005% d'utilisateurs visitent seulement un grand nombre de News via `read.php` sans visiter autres pages.

En utilisant des algorithmes d'extraction de motifs séquentiels, nous pouvons trouver le comportement le plus général décrit en (1) car le seuil de fréquence est élevé. Par contre, il est beaucoup plus difficile de découvrir le comportement des utilisateurs pour (2), (3) et (4) pour les raisons suivantes :

1. La plupart des approches existantes pour extraire des motifs séquentiels ne considère pas les éléments manqués, ni les contradictions sémantiques entre éléments (e.g. entre `cat1.html` et `cat5.html` dans une séquence). Même s'il existe des approches basées sur des contraintes comme l'approche SPIRIT proposée par Garofalakis et al. (1999) qui pourraient découvrir des séquences comme celles de (2) et (3), il n'est pas possible de spécifier que "des catégories contredisent sémantiquement `cat1.html`". En effet les approches basées sur les contraintes ne peuvent spécifier que "les catégories différentes de `cat5.html`", i.e. toutes les séquences qui ne contiennent pas `cat5.html`.

2. Dans une approche basée sur les motifs séquentiels, les séquences correspondant aux cas (2), (3) et (4) sont incluses dans celle du cas (1). Le modèle de *motif séquentiel clos* proposé par Yan et al. (2003) peut indiquer la valeur du support de chaque sous-séquence d'une séquence fréquente, donc l'existence des séquences pour (2), (3) et (4) peut être obtenue en comparant les valeurs du support de chaque sous-séquences d'une séquence fréquente. Cependant il n'est pas possible d'identifier telles séquences uniquement par leur valeur de support.

Dans cet article, nous proposons WebUser (**Web U**nexpected **se**quence **r**ules), une nouvelle approche d'extraction de comportements inattendus par rapport aux comportements connus via les données d'accès à un serveur Web. Le reste de cet article est organisé par la manière suivante. La section 2 présente les définitions préliminaires requises par notre approche, incluant

les définitions formelles de l'extraction de motifs séquentiels et la formalisation de la notion de session dans des logs d'accès. Dans la section 3, nous revenons sur la notion de comportement inattendu en spécifiant les notions de croyance, séquence inattendue, motif séquentiel inattendu et règle d'implication séquentielle inattendue. Nous proposons également notre approche WebUser ainsi que les algorithmes associés. La section 4 décrit les expérimentations menées sur des logs d'accès de différents serveurs Web. Les travaux liés sont présentés dans la section 5 et nous concluons cet article dans la section 6.

2 Définitions préliminaires

Dans cette section, nous introduisons les définitions formelles de l'extraction de motifs séquentiels et nous proposons une formalisation de la notion de sessions dans des logs d'accès.

2.1 Motifs séquentiels

Le problème de la recherche de motifs séquentiels dans une base de données a initialement été introduit par Agrawal et Srikant (1995) et est défini de la manière suivante.

Définition 1. *Soit un ensemble d'attributs $R = \{i_1, i_2, \ldots, i_n\}$, où chaque attribut correspond à un* item. *Un* itemset *est un ensemble d'items triés par l'ordre lexical, noté $I = (i_1 i_2 \ldots i_m)$ tel que $I \subseteq R$. Une* séquence *est une liste ordonnée d'itemsets, notée $s = I_1 I_2 \cdots I_k$.* □

Exemple 1. *Soit la séquence $s = (a)(bc)(d)$. Cette dernière peut être interprétée de la manière suivante "le client a acheté l'item a, puis en même temps les items b et c, et enfin l'item d".* □

Définition 2. *Soient deux séquences $s = I_1 I_2 \ldots I_m$ et $s' = I'_1 I'_2 \ldots I'_n$, s'il existe des entiers $1 \leq i_1 < i_2 < \ldots < i_m \leq n$ tels que $I_1 \subseteq I'_{i_1}, I_2 \subseteq I'_{i_2}, \ldots, I_m \subseteq I'_{i_m}$, alors on dit que la séquence s est une* sous-séquence *de la séquence s', notée comme $s \sqsubseteq s'$. Si $s \sqsubseteq s'$, on dit que s* est incluse dans *s', ou également s'* supporte *s. Si une séquence s n'est pas incluse dans aucune autres séquences, alors cette séquence s est une séquence* maximale. □

Exemple 2. *La séquence $s_1 = (a)(b)$ est incluse dans la séquence $s_2 = (a)(d)(bc)$ car $(a) \subseteq (a)$ et $(b) \subseteq (bc)$. En revanche, s_1 n'est pas incluse dans $s_3 = (ab)(d)$.* □

Définition 3. *Soit une base de données de séquences D, le* support *ou la* fréquence *d'une séquence s, noté $\sigma(s, D)$, est le nombre total de séquences dans D qui supportent s. Étant donné un seuil de fréquence minimal spécifié par l'utilisateur, noté min_supp, une séquence s est* fréquente *si $\sigma(s, D) \geq min\_supp$ et un* motif séquentiel *est une séquence fréquente maximale.* □

Généralement, dans le cas des données d'usage du Web, en considérant que les items peuvent correspondre à des URL, l'utilisation des motifs séquentiels permet d'extraire les comportements fréquents des utilisateurs sur un site.

2.2 Prise en compte des sessions

Dans le reste de l'article, nous considérons que les logs manipulés respectent le format NCSA Common Logfile Format (CLF) NCSA HTTPd Development Team (1995), qui est

utilisé par presque tous les serveurs Web (e.g. Apache http Server et Microsoft Internet Information Services). Le format CLF est organisé comme suit :

```
remotehost rfc931 authuser [date] "request" status bytes
```

Un log est donc un fichier texte où chaque ligne représente un accès réalisé par une machine distante. La figure 1(a) illustre des entrées, au format CLF, d'un log Apache HTTP Server. Des champs supplémentaires peuvent bien entendu exister comme `referer` et `user-agent` comme l'illustre la figure 1(b).

```
146.19.33.138 - - [11/Jan/2008:17:40:00 +0100] "GET /~li/ HTTP/1.1" 200 5480
146.19.33.138 - - [11/Jan/2008:17:40:00 +0100] "GET /~li/deepred.css HTTP/1.1" 304 -
146.19.33.138 - - [11/Jan/2008:17:40:27 +0100] "GET /~li/TP/TP07.htm HTTP/1.1" 200 2599
146.19.33.138 - - [11/Jan/2008:17:40:32 +0100] "GET /~li/TP/create.sql HTTP/1.1" 200 1376
146.19.33.138 - - [11/Jan/2008:17:49:21 +0100] "GET /~li/TP/TP07.pdf HTTP/1.1" 200 111134
```

(a)

```
146.19.33.138 - - [11/Jan/2008:18:27:35 +0100] "GET /~li/ HTTP/1.1" 200 1436 "-" "Mozilla
/5.0 (Macintosh; U; Intel Mac OS X; fr-fr) AppleWebKit/523.10.6 (KHTML, like Gecko) Versi
on/3.0.4 Safari/523.10.6"
146.19.33.138 - - [11/Jan/2008:18:29:38 +0100] "GET /~li/doc/ HTTP/1.1" 200 854 "http://
www.lgi2p.ema.fr/~li/" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X; fr-fr) AppleWebKit/523
.10.6 (KHTML, like Gecko) Version/3.0.4 Safari/523.10.6"
```

(b)

FIG. 1 – *(a) Un exemple de log au format CLF. (b) Un exemple de log étendu.*

Selon les définitions d'item, d'itemset et de séquence présentées dans la section précédente, nous proposons à présent la notion de session qui représente les activités liées à une session d'un utilisateur sur un serveur Web.

Définition 4. *Soit L un ensemble de logs et $l \in L$ une entrée dans le log. Une session $s \sqsubseteq L$ est une séquence définie par :*

$$s = (ip_s, S_0^s)(l_1^s.url, S_1^s) \dots (l_n^s.url, S_n^s),$$

telle que pour $1 \leq i \leq n$, $l_i^s.url$ est l'URL demandée par l'adresse IP ip_s, et pour tous $1 \leq i < j \leq n$, on a $l_i^s.time < l_j^s.time$, où $l_i^s.time$ et $l_j^s.time$ correspondent aux temps des requêtes de l_i^s et l_j^s. S_0^s est un ensemble d'items qui contiennent des propriétés pour la session s. $S_1^s \dots S_n^s$ sont les items qui contiennent des propriétés pour chaque entrée $l_1^s \dots l_n^s$. □

L'ensemble S_0^s peut être vide ou contenir des information sur l'adresse IP, la date, l'agent, etc., pour réduire la répétition des items. Les ensembles $S_1^s \dots S_n^s$ peuvent également être vides ou contenir les requêtes http de chaque entrée dans le log. La transformation en session est illustrée dans la figure 2, où #S correspond à l'identification de chaque séquence et #T indique l'identification de chaque transaction dans une séquence.

Dans la base de la figure 2, avec un support minimum de 0.5, nous pouvons trouver le motif séquentiel $(11, 15)(35)(52)$, qui correspond à la séquence d'accès suivante :

$$(\texttt{146.19.33.*}, \texttt{17h})(\texttt{TP07.html})(\texttt{TP07.pdf})$$

Cette séquence s'interprète de la manière suivante : à 17h, des utilisateurs qui viennent de l'adresse `146.19.33.*`, ont visité la ressource `TP07.html` et enfin la ressource `TP07.pdf`.

Session	No	IP/URL	Informations supplémentaires	#S	#T	Items
1	0	146.19.33.*	17h	1	1	11, 15
1	1	/~li/		1	2	21
1	2	/~li/deepred.css		1	3	22
1	3	/~li/TPBD/TP07.html		1	4	35
1	4	/~li/TPBD/create.sql		1	5	51
1	5	/~li/TPBD/TP07.pdf		1	6	52
2	0	146.19.33.*	17h	2	1	11, 15
2	1	/~li/TPBD/TP07.html		2	2	35
2	2	/~li/TPBD/TP07.pdf		2	3	52
2	3	/index.php	page=2	2	4	25, 59

Item	Contenu	Item	Contenu	Item	Contenu
11	146.19.3.*	15	17h	21	/~li/
22	/~li/deepred.css	25	/index.php	35	/~li/TPBD/TP07.html
51	/~li/TPBD/create.sql	52	/~li/TPBD/TP07.pdf	59	page=2

FIG. 2 – *Un exemple de sessions.*

3 Découverte de comportements inattendus

Dans cette section, nous proposons une approche, dirigée par les croyances, pour extraire des comportements inattendus dans une base de données de sessions.

3.1 Croyances et comportements inattendus

Dans la suite de cet article, nous considérons que les comportements connus sont représentés via des croyances. Cependant, avant de les définir formellement, nous introduisons quelques notions supplémentaires sur les séquences.

La *longueur* d'une séquence s est le nombre d'itemsets contenus dans la séquence et est notée $|s|$. La *concaténation* de séquences est notée $s_1 \cdot s_2$, ainsi nous avons $|s_1 \cdot s_2| = |s_1| + |s_2|$. Pour une séquence s, nous notons $s^\top$ le premier itemset et $s_\perp$ le dernier itemset. Ainsi, nous notons $s \sqsubseteq^\top s'$ si $s^\top \subseteq s'^\top$, $s \sqsubseteq_\perp s'$ si $s_\perp \subseteq s'_\perp$ et $s \sqsubseteq_\perp^\top s'$ si $s^\top \subseteq s'^\top \wedge s_\perp \subseteq s'_\perp$. Enfin, nous notons $s \sqsubseteq_c s'$ le fait que s est une *sous-séquence consécutive* de s'. Par exemple, nous avons $(a)(b)(c) \sqsubseteq_c (b)(\underline{a})(a, \underline{b})(\underline{c})(d)$, mais $(a)(b)(c) \not\sqsubseteq_c (b)(\underline{a})(a, \underline{b})(d)(\underline{c})(d)$.

Soit une séquence s telle que $s_1 \cdot s_2 \sqsubseteq s$, la *relation d'occurrence* est une contrainte $\mapsto^{\langle \mathbf{op}, n \rangle}$ sur des occurrences de s_1 et s_2 dans s, où $\mathbf{op} \in \{\neq, =, \leq, \geq\}$ et $n \in \mathbb{N}$. Nous notons $|s'| \models \langle \mathbf{op}, n \rangle$ si la longueur de la séquence s' satisfait la contrainte $\langle \mathbf{op}, n \rangle$. Ainsi, la relation $s_1 \mapsto^{\langle \mathbf{op}, n \rangle} s_2$ décrit $s_1 \cdot s' \cdot s_2 \sqsubseteq_c s$, où $|s'| \models \langle \mathbf{op}, n \rangle$. En outre, nous avons $\langle \leq, 0 \rangle$ implique $\langle =, 0 \rangle$. Par simplification, $s_1 \mapsto^{\langle \geq, 0 \rangle} s_2$ est noté $s_1 \mapsto^* s_2$, et $s_1 \mapsto^{\langle =, 0 \rangle} s_2$ est noté $s_1 \mapsto s_2$.

Exemple 3. *Nous avons* $|(a)(b)(c)| \models \langle \geq, 2 \rangle$, $|(a)(b)| \not\models \langle >, 2 \rangle$. *Nous avons également* $(a)(b)(c)(a)(b)(c)$ *satisfait* $(a)(b) \mapsto (c)$ *et* $(a)(b) \mapsto^{\langle \leq, 3 \rangle} (c)$. □

Définition 5. *Une* règle d'implication séquentielle, *notée* $s_\alpha \rightarrow s_\beta$, *où* s_α *et* s_β *sont deux séquences représente le fait que dans une séquence* s, *l'occurrence de la séquence* s_α *implique l'occurrence de la séquence* s_β. *Plus formellement,* $s_\alpha \sqsubseteq s \Rightarrow s_\alpha \cdot s_\beta \sqsubseteq s$. □

Selon la définition de règle d'implication séquentielle, deux contraintes supplémentaires, la contrainte d'occurrence et la contrainte de sémantique, peuvent être y ajoutées. La contrainte d'occurrence est une contrainte sur la relation d'occurrence, notée $s_\alpha \rightarrow^{\langle \mathbf{op}, n \rangle} s_\beta$, qui spécifie la relation entre s_α et s_β telle que $s_\alpha \sqsubseteq s \Rightarrow s_\alpha \cdot s' \cdot s_\beta \sqsubseteq_c s$, où $|s'| \models \langle \mathbf{op}, n \rangle$, c'est-à-dire que

s_α doit être suivie de s_β avec une intervalle en respectant $\langle \mathbf{op}, n \rangle$. La contrainte de sémantique est une contrainte sur la sémantique de séquences qui spécifie la contradiction sémantique entre deux séquences. Cette contradiction est déterminée par un prédicat $q(s_1, s_2)$ de deux séquences s_1 et s_2 : si s_1 contredit sémantiquement s_2, alors q retourne 1, sinon q retourne 0. Si $q(s_1, s_2) = 1$, on note alors $s_1 \not\approx_{sem} s_2$. Le prédicat q peut fonctionner par calculer la distance sémantique du concept associé avec les items contenant dans s_1 et s_2, ou simplement par examiner s_1 et s_2 dans une liste des séquences définie par des experts du domaine. Dans cet article, nous utilisons la dernière manière pour déterminer la contradiction sémantique entre des séquences, et de calculer la distance sémantique du concept est inclus dans nos perspectives de la recherche future. Ainsi, si on considère la contradiction sémantiquement de la séquence s_β, s'il existe une séquence s_γ telle que $s_\beta \not\approx_{sem} s_\gamma$, alors la règle $s_\alpha \rightarrow^{\langle \mathbf{op}, n \rangle} s_\beta$ décrit l'implication $s_\alpha \sqsubseteq s \Rightarrow s_\alpha \cdot s' \cdot s_\gamma \not\sqsubseteq s$, où $|s'| \models \langle \mathbf{op}, n \rangle$, c'est-à-dire que s_α ne doit pas être suivie de s_γ avec une intervalle en respectant $\langle \mathbf{op}, n \rangle$. Via ces contraintes, nous définissons la croyance de la manière suivante :

Définition 6. *Une* croyance *correspond à une règle d'implication séquentielle $s_\alpha \rightarrow^\tau s_\beta$ avec une contrainte d'occurrence $\tau = \langle \mathbf{op}, n \rangle$ et une contrainte sémantique $s_\beta \not\approx_{sem} s_\gamma$. Elle est notée $[s_\alpha; s_\beta; s_\gamma; \tau]$. Pour une séquence s nous avons : $s_\alpha \sqsubseteq s$ implique $s_\alpha \cdot s' \cdot s_\beta \sqsubseteq_c s$ mais $s_\alpha \cdot s' \cdot s_\gamma \not\sqsubseteq_c s$ où $|s'| \models \tau$.* □

Exemple 4. *Considérons le site Web de News présenté dans la section 1. Supposons que la plupart des utilisateurs visite la page* `index.html` *puis trois News via la page* `read.php`*, et enfin la page* `cat1.html`*. Ceci peut être représenté par la règle suivante :* (`index.html`) $\rightarrow^\tau$ (`cat1.html`) *avec comme contrainte d'occurrence $\tau = \langle \geq, 3 \rangle$. Si nous savons, à présent, que des visites de la page* `cat5.html` *ne doivent pas se faire trop tôt avant la visite de la page* `cat1.html` *(dans ce cas là, on considère que* `cat5.html` *est une contradiction sémantique de* `cat1.html` *puisque* `cat1.html` *ne doit pas être remplacée par* `cat5.html`*), nous pouvons ajouter la contrainte sémantique suivante :* (`cat1.html`) $\not\approx_{sem}$ (`cat5.html`)*. Finalement, nous obtenons la croyance suivante :*

$$[(\texttt{index.html}); (\texttt{cat1.html}); (\texttt{cat5.html}); \langle \geq, 3 \rangle]$$

sur les comportements des utilisateurs. □

Selon la contrainte d'occurrence et la contrainte sémantique utilisée, nous proposons trois catégories de séquences inattendues.

Définition 7. *Soient une croyance $b = [s_\alpha; s_\beta; s_\gamma; *]$ et une séquence s. Si $s_\alpha \sqsubseteq s$ et s'il n'existe pas de séquences s_β et s_γ telles que $s_\alpha \mapsto^* s_\beta \sqsubseteq s$ ou $s_\alpha \mapsto^* s_\gamma \sqsubseteq s$, alors la séquence s est une séquence α-inattendue par rapport à la croyance b.* □

Une croyance avec la contrainte d'occurrence $\tau = *$ affirme que s_β doit apparaître après l'occurrence de s_α dans une séquence s. La séquence s viole la contrainte $\tau = *$ si et seulement si $s_\alpha \sqsubseteq s \wedge s_\alpha \cdot s_\beta \sqsubseteq s$. Notons que pour ne pas mélanger les séquences inattendues à cause de l'occurrence ou de la sémantique, s_γ ne doit pas apparaître après l'occurrence de s_α dans une séquence α-inattendue. Par exemple, avec la croyance suivante

$$[(\texttt{index.html})(\texttt{read.php}); (\texttt{index.html}); \emptyset; *]$$

nous pourrions trouver des utilisateurs qui ne retournent jamais à la page `index.html` après avoir lu des News.

Définition 8. *Soient une croyance $b = [s_\alpha; s_\beta; s_\gamma; \tau]$ ($\tau \neq *$) et une séquence s. Si $s_\alpha \mapsto^* s_\beta \sqsubseteq s$ et s'il n'existe pas de séquence s' telle que $|s'| \models \tau$ et $s_\alpha \cdot s' \cdot s_\beta \sqsubseteq_c s$, alors la séquence s est une séquence β-inattendue par rapport à la croyance b.* □

Une séquence β-inattendue reflète le fait que la règle d'implication séquentielle est rompue car l'occurrence de s_β dans la séquence s viole la contrainte τ. Dans l'exemple précédent, il est attendu que la plupart des utilisateurs visite la page `cat1.html` après avoir lu au moins trois News depuis la page `index.html`. Il se peut cependant qu'il existe des utilisateurs qui visitent moins de trois News avant de quitter la page `index.html`. Via cette séquence β-inattendue par rapport à la croyance $[(\texttt{index.html}); (\texttt{cat1.html}); (\texttt{cat5.html}); \langle \geq, 3 \rangle]$, nous pourrions trouver que de tels comportements inattendus apparaissent, par exemple, dans des sites où le contenu n'est pas souvent mis à jour.

Définition 9. *Soient une croyance $b = [s_\alpha; s_\beta; s_\gamma; \tau]$ et une séquence s. Si $s_\alpha \mapsto^* s_\gamma \sqsubseteq s$ et s'il existe une séquence maximale s' telle que $|s'| \models \tau$ et $s_\alpha \cdot s' \cdot s_\gamma \sqsubseteq_c s$, alors la séquence s est une séquence γ-inattendue par rapport à la croyance b.* □

La notion de γ-inattendue concerne principalement la sémantique : l'occurrence de s_β est remplacée par sa contradiction sémantique s_γ dans la contrainte τ. Dans l'exemple précédent, nous considérons que les News de la page `cat1.html` sont sémantiquement différentes de celles de la page `cat5.html` (e.g. "les plus récentes" vs. "les anciennes", ou "politique" vs. "technologies"). Supposons à présent qu'un grand nombre d'utilisateurs visite les News sur `index.html` puis sur `cat1.html`, mais que peu d'utilisateurs visite `cat5.html` à la place de `cat1.html`, alors l'extraction d'un tel comportement inattendu pourrait faire émerger, par exemple, que même si nous savons que 60% des utilisateurs pendant la période 8h - 23h confirment le comportement attendu $(\texttt{index.html}) \mapsto^{\langle \geq, 3 \rangle} (\texttt{cat1.html})$ nous pouvons extraire que 80% des utilisateurs sur la période 23h - 8h confirment le comportement inattendu $(\texttt{index.html}) \mapsto^{\langle \geq, 3 \rangle} (\texttt{cat5.html})$. Ce type de comportements inattendus peut donc être utilisé pour extraire des périodes au cours desquelles l'usage fait du site Web est différent.

3.2 Motifs séquentiels et règles inattendus

Soient une croyance b et une séquence s, la séquence s est dite inattendue, notée $s \bowtie b$ si s viole la croyance b. Dans cette section, nous nous focalisons sur la partie inattendue et recherchons les causes qui sont amenées à avoir cet inattendu ainsi que les conséquences que cela engendre dans la base.

Définition 10. *Soient une croyance $b = [s_\alpha; s_\beta; s_\gamma; \tau]$ et une séquence inattendue $s \bowtie b$. La* caractéristique *de la séquence inattendue par rapport à la croyance b est une sous-séquence consécutive maximale u de s telle que : (1) si s est α-inattendue, nous avons $s_a \cdot u = s$ ($|s_a| \geq 0$) telle que $s_\alpha \not\sqsubseteq s_a$ et $s_\alpha \sqsubseteq^\top u$; (2) si s est β-inattendue, nous avons $s_a \cdot u \cdot s_c = s$ ($|s_a|, |s_c| \geq 0$) telle que $s_\alpha \not\sqsubseteq s_a$, $s_\alpha \sqsubseteq^\top u$ et $s_\beta \sqsubseteq_\perp u$; (3) si s est γ-inattendue, nous avons $s_a \cdot u \cdot s_c = s$ ($|s_a|, |s_c| \geq 0$) telle que $s_\alpha \not\sqsubseteq s_a$, $s_\alpha \sqsubseteq^\top u$ et $s_\gamma \sqsubseteq_\perp u$. La caractéristique est notée $u \models s \bowtie b$.* □

Étant donnée une séquence inattendue s par rapport à une croyance b, la caractéristique u est la partie de la séquence s qui cause l'inattendue $s \bowtie b$. A partir de cette caractéristique, nous recherchons à présent les raisons qui ont amenées à avoir cet inattendu via la notion de motifs séquentiels inattendus.

Définition 11. *Soient une croyance b et une base de données D. Soit $D_{(\triangleright\!\!\bowtie b)}$ un sous-ensemble de D constitué des séquences $s \in D$ telles que $s \bowtie b$. Soit $U_{(\triangleright\!\!\bowtie b)}$ l'ensemble des caractéristiques $u \models s \bowtie b$ de chaque séquence inattendue $s \in D_{(\triangleright\!\!\bowtie b)}$. Étant donné un seuil du support minimum min_supp, un* motif séquentiel inattendu *est une séquence maximale p dans l'ensemble de séquences $U_{(\triangleright\!\!\bowtie b)}$ telle que $\sigma(p, U_{(\triangleright\!\!\bowtie b)}) \geq min\_supp$.* □

(c) (f) (ab) (e) (f) (a) (a) (a) (a) (d) (bc) (a) (f)

(d) (b) (ab) (e) (f) (a) (a) (b) (a) (d) (bc) (g) (e)

(a) (x) (ab) (e) (f) (a) (e) (a) (e) (a) (e) (d) (bc) (x) (y)

(a) (c) (a) (b) (e) (f) (b) (b) (b) (c) (b) (ab) (c) (x)

(x) (a) (c) (y) (e) (b) (f) (b) (b) (c) (a) (ab) (d) (e) (y)

FIG. 3 – *Un exemple de séquences inattendues.*

Nous pouvons remarquer que, par construction, dans l'ensemble $U_{(\triangleright\!\!\bowtie b)}$ de caractéristiques le support de la partie inattendue est forcément de 100%. Par exemple pour une β-inattendue, le support de $s_\alpha \cdot s_\beta$ est de 100% car $s_\alpha \cdot s_\beta \sqsubseteq u$ pour chaque caractéristique $u \in U_{(\triangleright\!\!\bowtie b)}$. Aussi, pour rechercher les motifs inattendus, nous ne considérons pas les sous séquences s_α, s_β ou s_γ dans l'ensemble des caractéristiques.

Exemple 5. *Soit une croyance $b = [(e)(f); (d); (c); \langle \leq, 3 \rangle]$ où les étiquettes $a, b, c, \ldots$ représentent des items dans des données d'usages. Les séquences décrites dans la figure 3 sont β-inattendues et γ-inattendues. Dans les caractéristiques de β-inattendue, nous trouvons par exemple la séquence maximale $(a)(a)(a)$ qui est un motif séquentiel inattendu dont la présence indique que la séquence $(e)(f)$ donne le comportement inattendu : $(e)(f) \mapsto^{\langle >, 3 \rangle} (d)$. Dans les caractéristiques de γ-inattendue, nous trouvons le motif séquentiel inattendu $(b)(b)(b)$ dont la présence indique que la séquence $(e)(f)$ donne le comportement inattendu : $(e)(f) \mapsto^{\langle \leq, 3 \rangle} (c)$.* □

Définition 12. *Soient une séquence inattendue s et sa caractéristique u. La séquence s peut être représentée par $s = s_a \cdot u \cdot s_c$, où $|s_a|, |s_c| \geq 0$ (nous avons $|s_c| \equiv 0$ pour α-inattendue). La séquence s_a est appelée* séquence antécédente *d'inattendue et la séquence s_c est appelée* séquence conséquente *d'inattendue.* □

Définition 13. *Soient une croyance b et une base de données D. Soit $D^a_{(\triangleright\!\!\bowtie b)}$ un sous-ensemble de D constitué des séquences antécédentes s_a dans des séquences $s \in D$ telles que $s \bowtie b$. Une* règle antécédente *de l'inattendue $s \bowtie b$ est une règle $a \rightarrow (\bowtie b)$ où a est un motif séquentiel dans l'ensemble des séquences $D^a_{(\triangleright\!\!\bowtie b)}$.* □

Soient une croyance b et une base de données D. Le support d'une règle antécédente, noté $\sigma(a \rightarrow (\bowtie b), D)$, est le nombre total de séquences $s \in D$ telles que $s \bowtie b$. En d'autres termes, nous avons :

$$\sigma(a \rightarrow (\bowtie b), D) = |\{s \mid (s \bowtie b) \wedge (s \in D)\}| .$$

La confiance d'une règle antécédente, notée $\delta(a \rightarrow (\bowtie b), D)$, est le ratio de la séquence antécédente a dans l'ensemble $D^a_{(\bowtie b)}$ et dans la base D. En d'autres termes, nous avons :

$$\delta(a \rightarrow (\bowtie b), D) = \frac{\left|\{s \mid (s_a \sqsubseteq s) \wedge (s \in D^a_{(\bowtie b)})\}\right|}{|\{s \mid (s_a \sqsubseteq s) \wedge (s \in D)\}|}.$$

Les règles antécédentes d'inattendue reflètent les raisons qui ont amenées à avoir un inattendu. Supposons que, dans l'exemple 4, 80% des utilisateurs pendant la période 23h - 8h confirment le comportement inattendu (`index.html`) $\mapsto^{\langle \geq, 3 \rangle}$ (`cat5.html`), alors une règle antécédente peut être :

$$(\texttt{23h-08h}) \rightarrow (\texttt{index.html}) \mapsto^{\langle <, 3 \rangle} (\texttt{cat5.html}).$$

Pour cette règle, la confiance est de 80% et le support est le nombre de séquences de session qui supportent le comportement (`index.html`) $\mapsto^{\langle <, 3 \rangle}$ (`cat5.html`) dans la base de données entière.

Définition 14. *Soient une croyance b et une base de données D. Soit $D^c_{(\bowtie b)}$ un sous-ensemble de D constitué par les séquences conséquentes s_c dans des séquences $s \in D$ telles que $s \bowtie b$. Une* règle conséquente *de l'inattendue $s \bowtie b$ est une règle $(\bowtie b) \rightarrow c$ où c est un motif séquentiel dans l'ensemble des séquences $D^c_{(\bowtie b)}$.* □

Soient une croyance b et une base de données D. Le support d'une règle conséquente, noté $\sigma((\bowtie b) \rightarrow c, D)$, est le nombre total de séquences $s \in D^c_{(\bowtie b)}$ qui supportent la séquence conséquente c. En d'autres termes, nous avons :

$$\sigma((\bowtie b) \rightarrow c, D) = \left|\{s \mid (c \in s) \wedge (s \in D^c_{(\bowtie b)})\}\right| ;$$

La confiance d'une règle conséquente, noté $\delta((\bowtie b) \rightarrow c, D)$, est le ratio du support de la séquence conséquente c dans l'ensemble $D^c_{(\bowtie b)}$ et du nombre de séquences inattendues $s \bowtie b$ dans la base D. En d'autres termes, nous avons :

$$\delta((\bowtie b) \rightarrow c, D) = \frac{\left|\{s \mid (s_c \sqsubseteq s) \wedge (s \in D^c_{(\bowtie b)})\}\right|}{|\{s \mid (s \bowtie b) \wedge (s \in D)\}|}.$$

Les règles conséquentes d'inattendue reflètent les conséquences d'un inattendu. Dans l'exemple 4, considérons que 60% des utilisateurs qui accèdent au site pendant la période 23h - 8h confirment le comportement inattendu (`index.html`) $\mapsto^{\langle \geq, 3 \rangle}$ (`cat5.html`) puis visitent la page `ad3.php`, alors une règle conséquente peut être :

$$(\texttt{index.html}) \mapsto^{\langle <, 3 \rangle} (\texttt{cat5.html}) \rightarrow (\texttt{ad3.php}),$$

avec une confiance de 60%. Considérons, à présent, via les motifs séquentiels inattendus, que dans l'ensemble de caractéristiques de cet inattendu, nous ayons une sous-séquence fréquente (`read.php`) (`cat3.html`) avec un support de 90%. L'interprétation des comportements inattendus peut alors être le suivant :

"Entre 23h et 8h, 80% des utilisateurs ne respectent pas le comportement connu : visiter les News de la catégorie 5 après avoir lu trois News les plus récentes sur la page d'accueil. En effet, 90% d'entre eux visitent les News de la catégorie 3 au lieu de la catégorie 5, et 60% d'entre eux visitent les publicités en ligne proposées par le fournisseur 3."

3.3 L'approche WebUser

Dans cette section, nous décrivons l'approche WebUser pour découvrir des comportements inattendus dans des données d'usages. Notre approche respecte les principes généraux du WUM introduits par Srivastava et al. (2000) et illustrés dans la figure 4.

Dans un premier temps, nous convertissons les données d'usages en une base de sessions, notée D. La base de croyances, notée B, est construite à partir des comportements fréquents extraits via un algorithme d'extraction de motifs séquentiels sur la base de sessions mais également à partir de connaissances préalables sur les comportements attendus sur le site. L'objectif de l'algorithme WebUser-USE est d'extraire des séquences inattendues dans la base de sessions. Nous appliquons l'algorithme WebUser-USR qui extrait les motifs séquentiels inattendus, les séquences antécédentes et conséquentes et génère les règles associées.

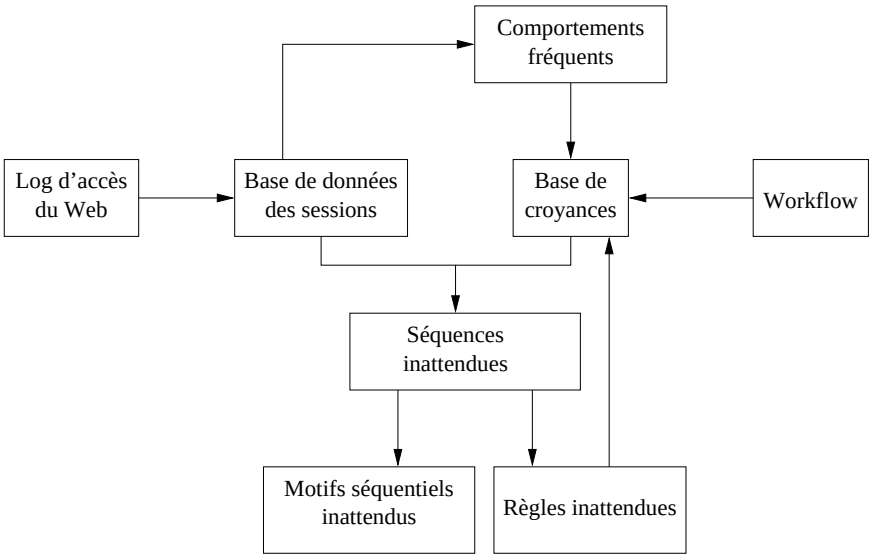

FIG. 4 – *Le processus général de l'approche WebUser.*

La base de croyances B est représentée via des arbres préfixés avec 3 blocs α, β and γ. Dans chaque bloc deux types d'arcs sont utilisés pour représenter les itemsets et les séquences et les blocs sont reliés par des liens pour représenter des croyances.

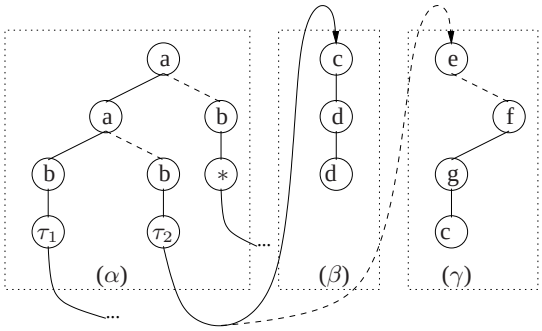

FIG. 5 – *Base de croyances en arbres préfixes.*

Considérons, par exemple, la croyance suivante $b = [s_\alpha; s_\beta; s_\gamma; \tau]$. Dans le bloc α, les séquences s_α de toutes les croyances sont organisées sous la forme d'un arbre préfixé. Dans la figure 5, (α) contient 3 séquences s_α : $(a)(a)(b)$, $(a)(ab)$ et (ab). Notons que pour améliorer l'efficacité de l'extraction, les τ sont dupliqués dans l'arbre. Par exemple, si $\tau_2 = \langle >, 5\rangle$, et que pour une séquence s, le parcours du chemin $(a) \mapsto (ab) \mapsto \tau_2$ arrive à b et que la longueur restante de la séquence s est inférieure à 5, la vérification de s avec le chemin s'arrête, i.e. le lien vers bloc β ne sera pas suivi. Par contre, le lien vers le bloc γ pourra être poursuivi s'il existe.

La procédure WebUser-USE (Algorithme 1) prend en entrée une base de séquences D et une base de croyances B puis retourne des ensembles de séquences inattendues ($D_{(\bowtie b)}$), de caractéristiques ($U_{(\bowtie b)}$), de séquences antécédentes ($D^a_{(\bowtie b)}$) et de séquences conséquentes ($D^c_{(\bowtie b)}$) pour chaque inattendu $(\bowtie\ b)$ de chaque croyance $b \in B$. Nous pouvons remarquer que contrairement à l'α-inattendue ou la γ-inattendue, dans le cas des β-inattendue, la fonction $seqinc(s, s_\beta, pos(b.s_\alpha), b.\tau)$ vérifie la présence de la séquence s_β dans la séquence s à partir de la fin de la séquence s_α en respectant la contrainte τ. Si s_α existe dans la séquence s et le reste de s satisfait la contrante τ et la longueur de s_β, l'algorithme continue le parcours de l'arbre de croyances et cherche la présence des toutes les s_β dans le bloc β et sortit s comme α-inattendue ou β-inattendue et les caractéristiques, séquences antécédentes et séquences conséquentes. L'algorithme puis parcourt le bloc γ pour vérifier si s est γ-inattendue et sortit les sous-séquences correspondantes.

Nous notons $P_{(\bowtie b)}$ l'ensemble des motifs séquentiels inattendus à partir de l'ensemble des caractéristiques, $F^a_{(\bowtie b)}$ l'ensemble de séquences antécédentes fréquentes et $F^c_{(\bowtie b)}$ l'ensemble de séquences conséquentes fréquentes par rapport à l'inattendue $(\bowtie\ b)$. Nous générons les règles inattendues avec la procédure WebUser-USR (Algorithme 2). L'algorithme génère les règles inattendues à partir des sous-ensembles des séquences extraits par l'algorithme WebUser-USE par rapport à la valeur de confiance minimum.

4 Expérimentations

De manière à évaluer notre approche, nous avons réalisé différentes expérimentations sur deux logs de serveurs Web d'une période de 3 mois. Le premier fichier BBS correspond à

Algorithm 1 Algorithme WebUser-USE

Require: Base de séquences D et base de croyances B

Ensure: Pour chaque inattendue ($\bowtie b$), ensemble $D_{(\bowtie b)}$ de séquences inattendues, $U_{(\bowtie b)}$ de caractéristiques, $D^a_{(\bowtie b)}$ de séquences antécédentes et $D^c_{(\bowtie b)}$ de séquences conséquentes

1: **for all** $s \in D$ **do**
2: **for all** $b \in B$ **do**
3: **if** $b.s_\alpha \sqsubseteq s$ **then**
4: /* vérifier la possibilité d'avoir des inattendues */
5: **if** $|s| - pos(s_\alpha) \models b.\tau$ **then**
6: /* découverte de α- ou γ-inattendue à partir de s_β */
7: **if** $T = follow\_link(b, \beta)$ **then**
8: **for all** $s_\beta \in T$ **do**
9: **if** $b.\tau == *$ **then**
10: /* découverte d'α-inattendue */
11: **if not** $seqinc(s, s_\beta, pos(b.s_\alpha), b.\tau)$ **then**
12: $output(s, D_{(\bowtie b)})$; /* enregistrer la séquence α-inattendue */
13: $output\_feature(s, U_{(\bowtie b)})$; /* enregistrer la future de l'inattendue */
14: $output\_seqante(s, D^a_{(\bowtie b)})$; /* enregistrer la séquence antécédente */
15: **end if**
16: **else**
17: /* découverte de β-inattendue */
18: **if** $seqinc(s, s_\beta, pos(b.s_\alpha), b.\tau)$ **then**
19: $output(s, D_{(\bowtie b)})$; /* enregistrer la séquence α-inattendue */
20: $output\_feature(s, U_{(\bowtie b)})$; /* enregistrer la future de l'inattendue */
21: $output\_seqante(s, D^a_{(\bowtie b)})$; /* enregistrer la séquence antécédente */
22: $output\_seqcons(s, D^c_{(\bowtie b)})$; /* enregistrer la séquence conséquente */
23: **end if**
24: **end if**
25: **end for**
26: **end if**
27: /* découverte de γ-inattendue à partir de s_γ */
28: **if** $T = follow\_link(b, \gamma)$ **then**
29: **for all** $s_\gamma \in T$ **do**
30: **if** $seqinc(s, s_\gamma, pos(b.s_\alpha), b.\tau)$ **then**
31: $output(s, D_{(\bowtie b)})$; /* enregistrer la séquence α-inattendue */
32: $output\_feature(s, U_{(\bowtie b)})$; /* enregistrer la future de l'inattendue */
33: $output\_seqante(s, D^a_{(\bowtie b)})$; /* enregistrer la séquence antécédente */
34: $output\_seqcons(s, D^c_{(\bowtie b)})$; /* enregistrer la séquence conséquente */
35: **end if**
36: **end for**
37: **end if**
38: **end if**
39: **end if**
40: **end for**
41: **end for**

Algorithm 2 Algorithme WebUSER-USR

Require: Ensemble $F^a_{(\bowtie b)}$ de séquences antécédentes fréquentes et $F^c_{(\bowtie b)}$ de séquences conséquentes fréquentes par rapport à l'inattendue $(\bowtie b)$, confiance minimum min_conf.

Ensure: Ensemble $R^a_{(\bowtie b)}$ de règles antécédentes et $R^c_{(\bowtie b)}$ de règles conséquent inattendues par rapport à l'inattendue $(\bowtie b)$

1: **for all** $s_a \in F^a_u$ **do**
2: **if** $\left| s_a \in D^a_{(\bowtie b)} \right| / |D| \geq min\_conf$ **then**
3: $R^a_{(\bowtie b)} = R^a_{(\bowtie b)} \cup \{s_a \rightarrow (\bowtie b)\}$; /* générer les règles antécédentes */
4: **end if**
5: **end for**
6: $output\_rules(R^a_{(\bowtie b)})$;
7: **for all** $s_c \in \mathcal{F}^c_{(\bowtie b)}$ **do**
8: **if** $\left| s_c \in D^c_{(\bowtie b)} \right| / \left| D_(\bowtie b) \right| \geq min\_conf$ **then**
9: $R^c_{(\bowtie b)} = R^c_{(\bowtie b)} \cup \{(\bowtie b) \rightarrow s_c\}$ /* générer les règles conséquentes */
10: **end if**
11: **end for**
12: $output\_rules(R^c_{(\bowtie b)})$;

un forum de discussion en PHP d'un fournisseur de jeux en ligne. Le second `WWW`, correspond au site Web d'un laboratoire qui fournit également l'hébergement des sites personnels d'enseignants-chercheurs.

Lors de nos expérimentations, nous divisons chaque fichier en trois fichiers d'une période d'un mois : `BBS-{1,2,3}` et `WWW-{1,2,3}`. Nous générons ensuite des sessions contenant l'information de la journée (du lundi au dimanche) et de l'heure (0H à 23H) dans la première entité d'une session. Si l'intervalle de temps entre deux accès de la même adresse IP d'un client est supérieur à 30 minutes, l'accès suivant correspond à une nouvelle session.

TAB. 1 – *Description des jeux de données utilisés.*

Nom du log	Sessions	Items distincts	Longueur moyenne
`BBS-1`	27 294	38 678	12,8934
`BBS-2`	47 868	42 052	20,3905
`BBS-3`	28 146	33 890	8,5762
`WWW-1`	6 534	8 436	6,3276
`WWW-2`	11 304	49 242	7,3905
`WWW-3`	28 400	50 312	9,5762

Pour chaque séquence de session, l'adresse IP du client est considérée comme le bloc d'adresses, tel que IP `146.19.33.138` est converti en `146.19.33.*`. Nous transférons uniquement les paramètres significatifs de requête HTTP aux items. Par exemple, pour `BBS`, la requête

```
/forumdisplay.php?f=2&sid=f2efeb85fcfd94ecbc2dba0f97b678a1
```

est considérée comme un itemset (f=2) correspondant à la visite du forum 2, et la requête

/viewtopic.php?t=57&sid=f2efeb85fcfd94ecbc2dba0f97b678a1

signifie un itemset (t=57) correspondant à la visite du sujet de discussion 57. Nous nous concentrons sur les accès des pages HTML statiques, les pages dynamiques et JavaScript, par conséquent tous les autres fichiers (e.g. .css, image et données) sont supprimés. Le tableau 1 détaille le nombre de séquences et d'items distincts, et la moyenne des longueurs des séquences.

Dans une première étape, nous appliquons tout d'abord un algorithme d'extraction de motifs séquentiels pour trouver les comportements fréquents dans BBS-{1,2,3} et WWW-{1,2,3} avec des supports minimum différents (c.f. Figure 6 et 7). Une analyse des résultats obtenus montre que, par exemple, dans les comportements fréquents découverts avec un support minimal 0.04 sur BBS-{1,2,3}, 149 motifs séquentiels découverts sont semblables (i.e. une similarité de plus de 40%) et que 197 motifs séquentiels sont semblables dans BBS-{2,3} (i.e. une similarité d'accès supérieure à 70%).

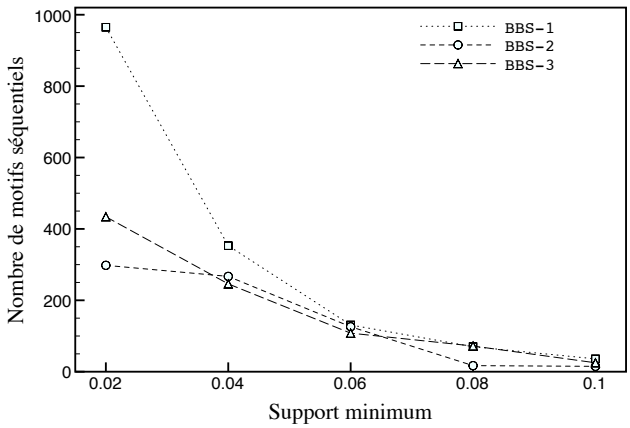

FIG. 6 – *Les comportements fréquents extraits dans* BBS.

Les bases de croyances sont ensuite construites manuellement à partir des connaissances de la structure du site et des comportements fréquents extraits lors de l'étape précédente. Pour BBS, nous construisons 5 croyances basées sur la structure du site puis considérons 5 croyances extraites des motifs séquentiels dans BBS-1. Par exemple, la croyance suivante correspond à un ordre "attendu" des parcours du forum (où t=2 correspond à la visite du sujet de discussion "règles d'utilisation" et t=5 correspond à "mode d'emploi", et les créateurs du site souhaitent que les utilisateurs puissent lire les règles d'utilisation avant lire le mode d'emploi) :

$$[(/); (\texttt{t=2})(\texttt{t=5}); (\texttt{t=5})(\texttt{t=2}); \langle =, 0\rangle].$$

Pour WWW, nous créons 10 croyances correspondant aux comportements les plus fréquents dans WWW-1. Par exemple, en fonction du menu de navigation de la page d'accueil du site Web, nous avons :

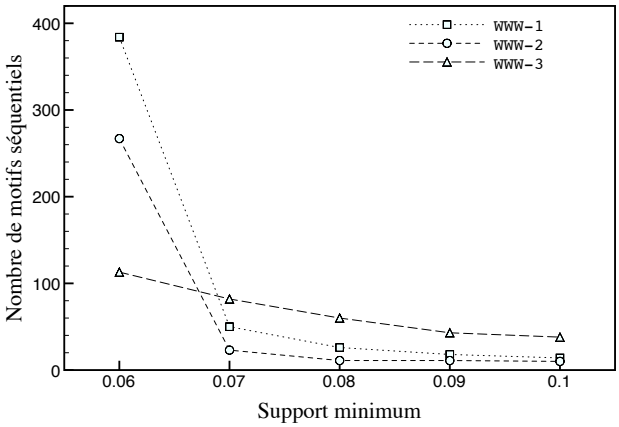

FIG. 7 – *Les comportements fréquents extraits dans* WWW.

$$[(\texttt{0018-04.html}); (\texttt{0019-27.html}); (\texttt{0018-04.html}); \langle \leq, 5 \rangle],$$

où `0018-04.html` correspond à la page d'index de la section "Research", `019-27.html` correspond à une sous-section de la section "Research" et `0018-04.html` correspond à une sous-section dans la section "Publications".

Les figures 8 et 9 indiquent le nombre de règles inattendues découvertes par notre approche WebUser. En comparaison avec les quantités de comportements fréquents extraits précédemment, notre approche génère moins de règles. Une analyse de la similarité montre que, avec une confiance minimale de 0.2, uniquement 3 règles inattendues sont similaires dans `BBS-{1,2,3}`, 4 règles similaires dans `BBS-{1,3}`, et enfin seulement 1 règle similaire dans `WWW-{2,3}`.

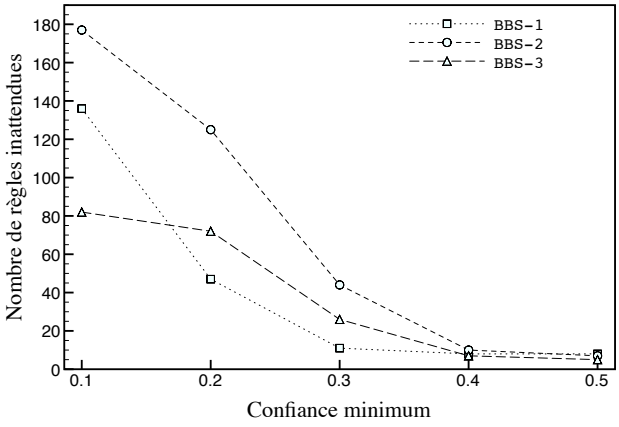

FIG. 8 – *Les règles inattendues découvertes par l'approche WebUser dans* BBS.

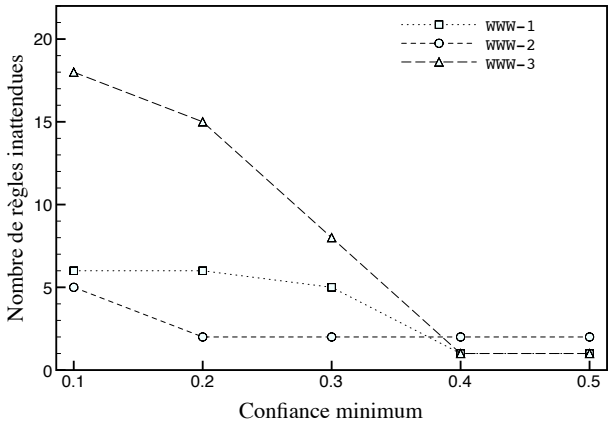

FIG. 9 – *Les règles inattendues découvertes par l'approche WebUser dans* WWW.

Nouvelles croyances peuvent être générées à partir des règles découvertes. Nous ajoutons dans les bases de croyances BBS-2 et WWW-2 10 règles inattendues découvertes dans BBS-1 et WWW-1 avec spécifier les contradictions sémantiques manuellement. Nous ajoutons également 10 règles inattendues (ou bien la totalité des règles si le nombre est inférieur à 10) découvertes dans (BBS-{1,2} et WWW-{1,2} dans la base de croyances pour BBS-3 et WWW-3. Le tableau 2 indique le nombre de règles obtenues en intégrant ces croyances, où les règles découvertes à partir de la base précédente ne sont pas incluses dans les résultats trouvés à partir de la base suivante.

TAB. 2 – *Le nombre de nouvelles règles inattendues en intégrant des comportements connus.*

min_conf	BBS-1	BBS-2	BBS-3	WWW-1	WWW-2	WWW-3
0.1	10 :136*	20 :24	30 :127	10 :6	16 :42	26 :34
0.2	10 :47	20 :18	30 :19	10 :6	16 :18	26 :25
0.3	10 :11	20 :16	30 :12	10 :5	15 :9	24 :12
0.4	10 :8	18 :9	27 :4	10 :1	11 :9	20 :10
0.5	10 :8	18 :6	24 :4	10 :1	11 :8	19 :10

*10 :136 indique que 136 règles inattendues sont trouvées par rapport à 10 croyances.

Par exemple, pour BBS, on souhaite que les visiteurs du forum 3 (les discussions sur un jeu "Mini F1 Match", noté J3) restent dans le forum, c'est-à-dire que l'itemset (f=3) est suivi de l'itemset (f=3) et $\tau = *$, mais on pense aussi que les joueurs du J3 peut-être ne pas jouer le jeu discuté dans le forum 6 (un petit jeu au sujet de la maison, noé J6), alors la contradiction sémantique peut être (f=3) $\not\simeq_{sem}$ (f=6). À partir de cette croyance, nous avons découvert les règles inattendues (dimanche) $\rightarrow \gamma$-inattendue et γ-inattendue $\rightarrow$ (f=7) dans la base BBS-1 (nous n'avons trouvé aucune séquence α-inattendue), où (f=7) correspond à un jeu de cartes, noté J7. De plus, comme une connaissance du côté du fournisseur du jeu, on sait que les

joueurs de J7 ne jouent pas souvent du jeu J5 (un jeu du billard), alors une nouvelle croyance peut être générer à partir de ces faits :

$$[(\texttt{dimanche})(\texttt{f=3}); (\texttt{f=7}); (\texttt{f=5}); *].$$

Cette croyance implique que à dimanche les joueurs visitent le forum du jeu J3 visiteront aussi le forum du jeu J7, par contre les joueurs qui jouent le jeu J7 ne jouent pas le jeu J5. Après l'avoir rajoutons à la base de croyances pour la découverte de `BBS-2`, nous avons découvert α`-inattendue` $\rightarrow$ (`f=4`) et γ`-inattendue` $\rightarrow$ (`f=5`). Ce n'est pas difficile à voir que la règle conséquente de la γ-inattendue est moins intéressante car elle signifie uniquement que les joueurs visitent plusieurs fois le forum du jeu J5.

Le principe de notre approche ne signifie pas que l'intégration des comportements inattendus les plus récents à la base de croyances va régulièrement affecter le nombre de règles inattendues dans les données des périodes suivantes. Cependant, le tableau 2 montre également que le nombre de règles inattendues augmente entre `WWW-1` et `WWW-{2,3}`. En fait, les données associées à `WWW` correspondent principalement à des données extraites de sites personnels où les accès sont très dépendants de la période (e.g. période d'un cours). Les croyances que nous avons spécifiées étant très générales, elles ne prennent donc pas en considération des cas trop spécifiques dans `WWW`.

5 Travaux antérieurs

Il existe actuellement de nombreux outils d'analyses de log pour extraire des informations par exemple Webalizer développé par Barrett (2006). Cependant, ces approches se basent principalement sur de simples requêtes (e.g. nombre de pages demandées, nombre de hits, ...) pour offrir au décideur des informations contenues dans les fichiers logs. Pour pallier ce problème de nombreuses approches se sont focalisées sur l'utilisation de techniques de fouille de données pour extraire des connaissances supplémentaires. Parmi les approches utilisées les techniques d'extraction de motifs séquentiels ont été particulièrement utilisées ces dernières années, par exemple Mobasher et al. (2002), Masseglia et al. (2003), Huang et al. (2006), Missaoui et al. (2007) et Masseglia et al. (2007)), car elles sont particulièrement adaptées aux données manipulées. Dans ce cadre, l'objectif essentiel de ces travaux est de transformer les données des fichiers logs pour pouvoir appliquer des algorithmes d'extraction de motifs.

A notre connaissance, notre approche WebUser est la première approche qui extrait des comportements inattendus dans des logs et qui tient compte de contraintes à la fois temporelles et sémantiques. En effet, dans ce cadre, nous proposons une mesure subjective à l'extraction de séquences. Présentées par Silberschatz et Tuzhilin (1995), les mesures d'intérêt pour la fouille de données peuvent généralement être classifiées en deux catégories : des mesures objectives ou des mesures subjectives. Les mesures objectives dépendent essentiellement de la structure des motifs extraits et de caractéristiques statistiques sur les données (e.g. le support et la confiance). Les mesures subjectives, par contre, tirent parti de connaissances sur le domaine ou sur l'expérience acquise via les utilisateurs (e.g. inattendue ou l'exigibilité).

Silberschatz et Tuzhilin (1995) introduisaient la notion d'inattendu à l'aide de croyances fortes et de croyances faibles. Une croyance forte est une croyance qui ne peut jamais être changée par de nouvelles évidences dans les données, et une contradiction de telle croyance implique une erreur de données. Par exemple, dans l'analyse de logs d'accès Web, l'erreur

"404 Not Found" peut être considérée comme une contradiction d'une croyance forte : "des ressources visitées doivent être disponibles". D'autre part, une croyance faible correspond à des contraintes sur les données mesurées par un degré qui peut être modifié via de nouvelles évidences contredisant cette croyance. L'intérêt de ces nouvelles évidences est mesuré par le changement du degré. Dans cet article, nous proposons une application des croyances faibles. Par exemple, dans un ensemble de données, nous savons que 90% d'utilisateurs visitent `cat1.html` et ensuite `cat5.html`. Il est donc possible de créer la croyance faible suivante : "l'accès de `cat1.html` implique l'accès de `cat5.html`" et son degré peut être défini par une fonction $\mu(0.9)$. Si dans un nouvel ensemble de données, il y a seulement 10% d'utilisateurs qui vérifient cette croyance, alors le changement de degré peut être donné par une fonction $\delta(0.9, 0.1)$.

La prise en compte de la notion d'inattendu dans le cadre de la fouille de données considère généralement les règles associations inattendues, par exemple l'approche de Padmanabhan et Tuzhilin (2006). Dans cette approche, les auteurs proposent la notion d'inattendu basée sur la sémantique des données. Par exemple, une règle d'association $A \rightarrow B$ est inattendue par rapport à la règle $X \rightarrow Y$ si : (1) B et Y sont sémantiquement opposés (on notera $BETY \models FAUX$) ; (2) le support et la confiance de la règle $A \cup X \rightarrow B$ sont suffisants ; (3) le support et la confiance de la règle $A \cup X \rightarrow Y$ ne sont pas suffisants.

Spiliopoulou (1999) proposaient d'extraire des motifs séquentiels inattendus à partir de logs d'accès Web. Cette approche crée des croyances à l'aide de la conjonction des fréquences de certains éléments dans une séquence : si une séquence fréquente ne respecte pas les fréquences d'éléments, alors elle est considérée comme inattendue. La découverte de règles séquentielles à partir de telles séquences inattendues est aussi proposée. Même si ces travaux considèrent des séquences inattendues, ils sont cependant différents de notre problématique dans la mesure ou la notion d'inattendu concerne des séquences fréquentes sur la base afin de trier les résultats obtenus. Notre objectif est différent car nous souhaitons extraire, à partir d'une base, toutes les séquences inattendues et générer également des règles elles-mêmes inattendues. Pour cela, nous considérons à la fois l'aspect inattendu par rapport à une connaissance du domaine (mesure subjective) et l'aspect "valide" au sens classique du support et de la confiance (mesures objectives).

6 Conclusion

Dans cet article, nous avons présenté l'approche WebUser pour découvrir des comportements inattendus dans des logs d'usages. Dans un premier temps, nous avons proposé une définition formelle de la notion de session et avons formalisé la notion de base de croyances sur des comportements par rapport aux contraintes sémantiques et temporelles. Ensuite, nous avons proposé trois types d'inattendus, les motifs séquentiels et règles inattendus associés.

Nous avons expérimenté notre approche WebUser sur des logs d'accès de différents types. Les résultats expérimentaux montrent que : (1) notre approche permet d'extraire des comportements inattendus même avec un support très faible ; (2) notre approche est capable de trouver des sessions inattendues qui sont incomplètes par rapport aux comportements prévus, i.e. des sous-séquences encore plus fréquentes que des séquences fréquentes prévues ; (3) la recherche de comportements inattendus dans des logs de manière incrémentale évite de redécouvrir des

comportements connus ; (4) les comportements inattendus dépendent fortement des comportements connus.

Nous souhaitons à présent extraire des comportements inattendus en introduisant la notion de hiérarchie et de hiérarchie floue. Nous souhaitons également étendre les contraintes en intégrant des contraintes floues et ainsi ajouter plus de flexibilité dans l'expression des contraintes temporelles.

Références

Agrawal, R. et R. Srikant (1995). Mining sequential patterns. In *ICDE*, pp. 3–14.

Barrett, B. L. (1997-2006). Webalizer. http ://www.mrunix.net/webalizer/.

Büchner, A. G. et M. D. Mulvenna (1998). Discovering internet marketing intelligence through online analytical web usage mining. *SIGMOD Record 27*(4), 54–61.

Eirinaki, M. et M. Vazirgiannis (2003). Web mining for web personalization. *ACM Trans. Internet Techn. 3*(1), 1–27.

Garofalakis, M. N., R. Rastogi, et K. Shim (1999). SPIRIT : Sequential pattern mining with regular expression constraints. In *VLDB*, pp. 223–234.

Huang, Y.-M., Y.-H. Kuo, J.-N. Chen, et Y.-L. Jeng (2006). NP-miner : A real-time recommendation algorithm by using web usage mining. *Knowl.-Based Syst. 19*(4), 272–286.

Masseglia, F., P. Poncelet, M. Teisseire, et A. Marascu (2007). Web usage mining : Extracting unexpected periods from web logs. In *DMKD*.

Masseglia, F., M. Teisseire, et P. Poncelet (2003). HDM : A client/server/engine architecture for real-time web usage mining. *Knowl. Inf. Syst. 5*(4), 439–465.

Missaoui, R., P. Valtchev, C. Djeraba, et M. Adda (2007). Toward recommendation based on ontology-powered web-usage mining. *IEEE Internet Computing 11*(4), 45–52.

Mobasher, B. (2007). Data mining for web personalization. In *The Adaptive Web*, pp. 90–135.

Mobasher, B., H. Dai, T. Luo, et M. Nakagawa (2002). Using sequential and non-sequential patterns in predictive web usage mining tasks. In *ICDM*, pp. 669–672.

NCSA HTTPd Development Team (1995). NCSA HTTPd Online Document : TransferLog Directive. http ://hoohoo.ncsa.uiuc.edu/docs/setup/httpd/TransferLog.html.

Padmanabhan, B. et A. Tuzhilin (2006). On characterization and discovery of minimal unexpected patterns in rule discovery. *IEEE Trans. Knowl. Data Eng. 18*(2), 202–216.

Silberschatz, A. et A. Tuzhilin (1995). On subjective measures of interestingness in knowledge discovery. In *KDD*, pp. 275–281.

Spiliopoulou, M. (1999). Managing interesting rules in sequence mining. In *PKDD*, pp. 554–560.

Spiliopoulou, M., C. Pohle, et L. Faulstich (1999). Improving the effectiveness of a web site with web usage mining. In *WEBKDD*, pp. 142–162.

Srivastava, J., R. Cooley, M. Deshpande, et P.-N. Tan (2000). Web usage mining : Discovery and applications of usage patterns from web data. *SIGKDD Explorations 1*(2), 12–23.

Yan, X., J. Han, et R. Afshar (2003). CloSpan : Mining closed sequential patterns in large databases. In *SDM*, pp. 166–177.

Summary

In recent years, Web usage mining is more and more concentrated on finding valuable user behaviors from Web navigation record data. Although the sequential pattern mining has been well adopted for finding frequent user behaviors, the decision makers will be more and more interested in unexpected behaviors that do not confirm the belief base on existing domain knowledge. In this paper, we present a belief-driven approach, WebUser, for discovering unexpected behaviors from Web access logs. Our experiments with the belief bases constructed from the explored user behaviors show that our approach is effective and useful to extract unexpected behaviors for improving the Web site structures and user experiences.