Découverte de motifs séquentiels et de règles inattendus

D. H. Li*, A. Laurent**, P. Poncelet*

*LGI2P - EMA, SITE EERIE
{haoyuan.li,pascal.poncelet}@ema.fr

**LIRMM - CNRS - Université Montpellier II
laurent@lirmm.fr

Résumé. Les travaux autour de l'extraction de motifs séquentiels se sont particulièrement focalisés sur la définition d'approches efficaces pour extraire, en fonction d'une fréquence d'apparition, des corrélations entre des éléments dans des séquences. Même si ce critère de fréquence est déterminant, le décideur est également de plus en plus intéressé par des connaissances qui sont représentatives d'un comportement inattendu dans ces données (erreurs dans les données, fraudes, nouvelles niches, ...). Dans cet article, nous introduisons le problème de la détection de motifs séquentiels inattendus par rapport aux croyances du domaine. Nous proposons l'approche USER dont l'objectif est d'extraire les motifs séquentiels et les règles inattendues dans une base de séquences.

1 Introduction

Pour faire face aux besoins des nouvelles applications (médicales, suivi de consommation, suivi des navigations sur un serveur Web, etc), de plus en plus de données sont stockées sous la forme de séquences. Pour traiter ces bases et en extraire des connaissances pertinentes, les motifs séquentiels ont été proposés Agrawal and Srikant (1995). Ils permettent, étant donnée une base de données de séquences, de trouver toutes les séquences maximales fréquentes au sens d'un support minimal défini par l'utilisateur.Si la découverte de corrélations dans les données séquentielles est primordiale pour le décideur, il n'en reste pourtant pas moins que certains problèmes ne peuvent être résolus par la recherche de tendances. De nouveaux motifs intéressent le décideur : les motifs inattendus qui contredisent les croyances acquises sur le domaine pour, par exemple, détecter des attaques sur un réseau.

Rappelons que notre objectif n'est pas de trouver les motifs rares, mais bien les motifs contredisant une connaissance, ce qui n'existe pas dans la littérature. La recherche de connaissance inattendue à partir d'une base de croyance a été introduite dans Silberschatz and Tuzhilin (1995) et Padmanabhan and Tuzhilin (2006) présentent une approche de découverte de règles d'association inattendues. Spiliopoulou (1999) propose un cadre basé sur la connaissance du domaine et des croyances pour trouver des règles séquentielles inattendues à partir de séquences fréquentes. Même si ces travaux considèrent des séquences inattendues, ils sont différents de notre problématique dans la mesure où la notion d'inattendue concerne des séquences fréquentes sur la base afin de trier les résultats obtenus. Notre objectif est d'extraire, à

partir d'une base, toutes les séquences inattendues et d'obtenir des règles elles mêmes inattendues.

Dans cet article, nous définissons donc la notion de base de croyances et de contradiction dans le contexte des séquences. Nous parlons alors de séquence inattendue, et nous introduisons les méthodes de découverte de telles séquences. Etant donné que les motifs séquentiels traditionnels ne mettent pas en évidence des règles du type "antécédent-conséquent", nous étendons la notion de séquences inattendues à celles de règles inattendues. Pour extraire ces règles à partir d'une base de données de séquences et d'une base de croyances, nous proposons l'approche USER (Unexpected Sequence Extracted Rules).

2 USER : Motifs séquentiels et règles inattendus

Soit un ensemble d'attributs distincts, on nomme **item** i un attribut de cet ensemble. Un **itemset** $\mathcal{I}$ est une collection non ordonnée d'items, notée $(i_1 i_2 \dots i_m)$. Une **séquence** s est une liste ordonnée d'itemsets, notée $\langle \mathcal{I}_1 \mathcal{I}_2 \dots \mathcal{I}_k \rangle$. Une **base de séquences** $\mathcal{D}$ est un ensemble de séquences (de taille potentiellement très grande).

Soient deux séquences $s = \langle \mathcal{I}_1 \mathcal{I}_2 \dots \mathcal{I}_m \rangle$ et $s' = \langle \mathcal{I}'_1 \mathcal{I}'_2 \dots \mathcal{I}'_n \rangle$, on dit que s est une **sous-séquence** de s', notée $s \sqsubseteq s'$ (s *est contenu dans* s') s'il existe des entiers $1 \leq i_1 < i_2 < \dots < i_m \leq n$ tels que $\mathcal{I}_1 \subseteq \mathcal{I}'_{i_1}, \mathcal{I}_2 \subseteq \mathcal{I}'_{i_2}, \dots, \mathcal{I}_m \subseteq \mathcal{I}'_{i_m}$. Dans un ensemble de séquences, si une séquence s n'est une sous-séquence d'aucune autre, elle est dite **maximale** ; sinon, si s est contenue dans s', on dit que s' **supporte** la séquence s. Soit une séquence $s = \langle \mathcal{I}_1 \mathcal{I}_2 \dots \mathcal{I}_k \rangle$, un **segment** $g \sqsubseteq s$ est une sous-séquence qui contient des itemsets contigus $\langle \mathcal{I}_i \mathcal{I}_{i+1} \dots \mathcal{I}_{i+n} \rangle$ avec $i \leq 1$ et $i + n \leq k$. Le **support** d'une séquence est défini comme la proportion de séquences dans $\mathcal{D}$ qui supportent cette séquence.

Nous définissons, dans cet article, la *longueur* d'une séquence comme le nombre d'itemsets qu'elle contient, noté $|s|$. Nous considérons également la séquence vide et la concaténation de séquences. Une *séquence vide* est notée $\emptyset$, avec $s = \emptyset \iff |s| = 0$. Soient deux séquences s_1 et s_2, la *concaténation* $s_1 \cdot s_2$ de ces deux séquences correspond à s_1 complétée par s_2 en fin de séquence. Nous avons alors $|s_1 \cdot s_2| = |s_1| + |s_2|$.

Pour simplifier les notations et la lecture, dans la suite de cet article, nous utilisons les majuscules $A, B, C \dots$ pour décrire des items, et la notation (ABC) pour désigner des itemsets. La notation $\langle (A)(AC)(BC) \rangle$ désigne une séquence (A puis A et C puis B et C).

Nous notons $\langle \mathbf{op}, n \rangle$ une contrainte sur la longueur de séquences, avec $\mathbf{op} \in \{\neq, =, <, \leq, >, \geq\}$ et $n \in \mathbb{N}$. La notation $|s'| \models \langle \mathbf{op}, n \rangle$ signifie que la longueur de la séquence satisfait $\langle \mathbf{op}, n \rangle$. Dans le cas où l'on a $\langle \mathbf{op}, n \rangle = \langle \geq, 0 \rangle$, on note $*$.

Soit une séquence s, avec s_1 et s_2 deux sous-séquences de s, i.e. $s_1, s_2 \sqsubseteq s$, telles que s_1 apparaisse avant s_2 dans s. On a donc $s = s_1 \cdot g \cdot s_2$. L'expression $s_1 \mapsto^{\langle \mathbf{op}, n \rangle} s_2$ indique que s_1 et s_2 apparaissent dans la séquence $s = s_1 \cdot g \cdot s_2$, avec g vérifie $\langle \mathbf{op}, n \rangle$. Dans le cas où $\langle \mathbf{op}, n \rangle = \langle =, 0 \rangle$ ($g = \emptyset$), nous notons $s_1 \mapsto s_2$ le fait que s_1 soit directement suivi de s_2 dans la séquence s. Nous avons alors : $\left\langle s_1 \mapsto^{\langle =, 0 \rangle} s_2 \right\rangle \equiv \langle s_1 \mapsto s_2 \rangle$. Dans le cas le moins contraint, l'écriture $s_1 \mapsto^* s_2$ désigne le fait que s_2 apparaît après s_1 dans s ($s = s_1 \cdot s' \cdot s_2$ sans contrainte sur s').

Une croyance représente une connaissance décrite sous la forme d'une relation de causalité temporelle entre des occurrences d'éléments dans une séquence.

Dans notre approche, nous utilisons des *règles* pour décrire les relations de causalité entre séquences. De manière similaire à Spiliopoulou (1999), nous notons s_α la prémisse et s_β la conclusion. Ceci signifie que $s_\alpha \Rightarrow s_\beta$ est satisfaite dans s, si l'occurrence de $s_\alpha \sqsubseteq s$ implique l'occurrence d'une sous-séquence $s_\beta \sqsubseteq s$ telle que $s_\alpha \cdot s_\beta \sqsubseteq s$.

Dans notre approche, ce sont les experts qui définissent les croyances à prendre en compte.

Définition 1 (Croyance)**.** *Une croyance b sur une séquence est un couple $(p, \mathcal{C})$ tel que : b : $(p, \mathcal{C})$ où $p : s_\alpha \Rightarrow s_\beta$, $\mathcal{C} : \{\tau, \eta\}$, $\tau : \langle \mathbf{op}\, n \rangle, \mathbf{op} \in \{\neq, =, <, \leq, >, \geq\}, n \in \mathbb{N}$, et η : $s_\beta \not\sim s_\gamma$.*

p et τ forment alors une règle $s_\alpha \mapsto^{\langle \mathbf{op}, n \rangle} s_\beta$ et η spécifie que l'occurrence de s_β ne peut pas être remplacée par une occurrence de s_γ. La croyance c est alors notée $[s_\alpha; s_\beta; s_\gamma; \tau]$. Dans le cas où $\tau = \langle \geq, 0 \rangle$, on note $*$.

Soit une croyance b, une séquence s est *inattendue* par rapport à b si s viole l'une des contraintes introduites par b. En nous appuyant sur ces contradictions possibles de contraintes, nous distinguons trois types de violation (caractères inattendus) : α-inattendu, β-inattendu, γ-inattendu.

Définition 2. *Soit une croyance $b = [s_\alpha; s_\beta; s_\gamma; *]$ et une séquence s. s est dite α-***inattendue** *par rapport à b si $s_\alpha \sqsubseteq s$ et il n'existe pas s_β, s_γ tels que $\langle s_\alpha \mapsto^* s_\beta \rangle \sqsubseteq s$ ou $\langle s_\alpha \mapsto^* s_\gamma \rangle \sqsubseteq s$. Soit une croyance $b = [s_\alpha; s_\beta; s_\gamma; \tau]$ où la contrainte τ est différente de $*$, et une séquence s. s est dite β-***inattendue** *par rapport à b si $\langle s_\alpha \mapsto^* s_\beta \rangle \sqsubseteq s$ et s'il n'existe pas s' tel que $|s'| \models \tau$ et $\langle s_\alpha \mapsto s' \mapsto s_\beta \rangle \sqsubseteq s$, Soit une croyance $b = [s_\alpha; s_\beta; s_\gamma; \tau]$ et une séquence s. s est dite γ-***inattendue** *par rapport à b si $\langle s_\alpha \mapsto^* s_\gamma \rangle \sqsubseteq s$ et il existe s' tel que $|s'| \models \tau$ et $\langle s_\alpha \mapsto s' \mapsto s_\gamma \rangle \sqsubseteq s$.*

Notons qu'une séquence inattendue peut être liée à deux contraintes violées. Ainsi une séquence peut être à la fois α- et γ-inattendue, ou être à la fois β- et γ-inattendue.

Il est à présent intéressant de découvrir les tendances au sein de ces séquences (par exemple pour caractériser une attaque ou une fraude). Nous utilisons pour cela le cadre des motifs séquentiels. À partir d'une base de données de séquences $\mathcal{D}$ et d'une base de croyances $\mathcal{B}$, nous utilisons le support pour déterminer à quel point une séquence inattendue s_u issue d'une séquence $s \in \mathcal{D}$ est fréquente au sens de la croyance $b \in \mathcal{B}$ et du type d'inattendu $u \in \{\alpha, \beta, \gamma\}$. On note D_u le sous-ensemble de D composé des séquences violant b pour le type d'inattendu u. On a alors : $supp(s_u) = \dfrac{|\{s \in \mathcal{D}_u | s_u \sqsubseteq s\}|}{|\mathcal{D}_u|}$.

Définition 3. *Soit une base de données $\mathcal{D}$, u un type d'inattendu de contrainte, et $\mathcal{D}_u$ le sous-ensemble de $\mathcal{D}$ contenant les séquences $s \sqsubseteq \mathcal{D}$ telles que s est une u-inattendu de croyance. Un* **motif séquentiel inattendu** *est une séquence maximale fréquente, c'est-à-dire dont le support est supérieur à un seuil fixé par l'utilisateur.*

Afin de mieux identifier les segments de séquence correspondant à la partie "souche" et aux parties de contradiction, nous introduisons la notion de *séquence inattendue bornée*.

Définition 4. *Soit une croyance $b = [s_\alpha; s_\beta; s_\gamma; *]$ et une séquence s telle que $s = g' \cdot s_\alpha \cdot g$ avec $s_\alpha \not\sqsubseteq g'$, $s_\beta \not\sqsubseteq g$, $s_\gamma \not\sqsubseteq g$ (s est α-inattendue pour b). La séquence α-***inattendue bornée** *est définie comme le segment $s_b = s_\alpha \cdot g$. Soit une croyance $b = [s_\alpha; s_\beta; s_\gamma; \tau]$ et une séquence*

*s telle que $s = g' \cdot s_\alpha \cdot g \cdot s_\beta \cdot g''$ où le segment g ne satisfait pas τ (s est β-inattendue pour b). La séquence β-***inattendue bornée*** est définie comme le segment $s_b = s_\alpha \cdot g \cdot s_\beta$. Soit une croyance $b = [s_\alpha; s_\beta; s_\gamma; \tau]$ et une séquence $s = g' \cdot s_\alpha \cdot g \cdot s_\gamma \cdot g''$ où le segment g satisfait τ (s est γ-inattendue pour b). La séquence γ-***inattendue bornée*** est définie comme le segment $s_b = s_\alpha \cdot g \cdot s_\gamma$.*

Une séquence inattendue s peut donc être représentée comme $s = g_a \cdot s_b \cdot g_c$ où s_b est une séquence bornée inattendue correspondant à la violation de contrainte de croyance, et g_a, g_c sont deux segments de s. Nous avons $|s_b| > 0$, $|g_a| \geq 0$ et $|g_c| \geq 0$. Le segment $g_a \sqsubseteq s$ est appelé *antécédent* et le segment $g_c \sqsubseteq s$ est appelé *conséquent*. Le support d'une séquence s_a contenu dans $\mathcal{D}_u^a$ et s_c contenu $\mathcal{D}_u^c$ est la fraction de séquences de $\mathcal{D}_u^a$ ou de $\mathcal{D}_u^c$ qui supportent s_a ou s_c, c'est-à-dire $supp(s_a) = \frac{|\{s \in \mathcal{D}_u^a | s_a \sqsubseteq s\}|}{|\mathcal{D}_u^a|}$ et $supp(s_c) = \frac{|\{s \in \mathcal{D}_u^c | s_c \sqsubseteq s\}|}{|\mathcal{D}_u^c|}$.

Une séquence maximale fréquente contenue dans $\mathcal{D}_u^a$ est une *séquence antécédent fréquente* et une séquence fréquente maximale contenue dans $\mathcal{D}_u^c$ est une *séquence fréquente conséquente*.

Définition 5 (Règle Antécédente et Règle Conséquente)**.** *Soit un ensemble $\mathcal{D}_u$ de séquences inattendues de type $u \in \{\alpha, \beta, \gamma\}$. Soit $\mathcal{D}_u^a$ l'ensemble de toutes les séquences antécédentes contenues dans $\mathcal{D}_u$ et s_a une séquence antécédente contenue dans $\mathcal{D}_u^a$ fréquente par rapport à un support défini par l'utilisateur σ_a. Une règle antécédente est une règle de la forme $s_a \Rightarrow u$. Soit $\mathcal{D}_u^c$ l'ensemble de toutes les séquences conséquentes contenues dans $\mathcal{D}_u$ et soit s_c la séquence conséquente contenue dans $\mathcal{D}_u^c$ fréquente par rapport au seuil σ_c fixé par l'utilisateur. Une règle conséquente est une règle de la forme $u \Rightarrow s_c$.*

Les règles antécédentes reflètent les éléments d'une séquence qui sont en amont d'une violation de contrainte d'une croyance donnée. Les règles conséquentes reflètent les éléments d'une séquence qui correspondent au caractère inattendu par rapport à une croyance.

Décrites par leur support, les règles antécédentes et conséquentes sont également décrites par leur *confiance*. Nous avons donc $supp(s_a \Rightarrow u) = supp(s_a)$ et $supp(u \Rightarrow s_c) = supp(s_c)$.

Définition 6 (Support et Confiance de règles)**.** *Soit une base de données de séquences $\mathcal{D}$ et un type d'inattendu u. Soit l'ensemble d'antécédents $\mathcal{D}_u^a$ et l'ensemble des conséquents $\mathcal{D}_u^c$. La valeur du support d'une règle $s_a \Rightarrow u$ équivaut à la valeur de support de s_a de même que la valeur de support de $u \Rightarrow s_c$ est égale à la valeur de support de s_c.*

$$conf(s_a \Rightarrow u) = \frac{|\{s \in \mathcal{D}_u^a | s_a \sqsubseteq s\}|}{|\{s \in \mathcal{D} | s_a \sqsubseteq s\}|}, conf(u \Rightarrow s_c) = \frac{|\{s \in \mathcal{D}_u^c | s_c \sqsubseteq s\}|}{|\{s \in \mathcal{D} | s \models u\}|}.$$

2.1 L'approche USER

Nous supposons connue une base de croyances et cherchons les comportements inattendus dans une base de données de séquences. L'approche décrite ici s'articule autour de deux phases. Dans la première phase, l'algorithme USE (Unexpected Sequence Extraction) extrait toutes les séquences inattendues pour chaque type de violation de contrainte et pour chaque croyance. Dans une seconde phase, l'algorithme USR (Unexpected Sequence Rules) trouve tous les motifs séquentiels inattendus et les règles associées à partir des séquences inattendues trouvées par USE, à partir de seuils de support/confidence définis a priori. Différentes

Algorithm 1 Algorithme USE

Input: Base de séquences $\mathcal{D}$ et base de croyances $\mathcal{B}$

Output: Ensemble $\mathcal{D}_u$ de séquences inattendues, $\mathcal{D}_u^a$ de séquences antécédentes et $\mathcal{D}_u^c$ de séquences conséquentes pour chaque type de violation u

for all $s \in \mathcal{D}$ **do**
  for all $b \in \mathcal{B}$ **do**
    /* α-inattendus */
    for all $u_\alpha \vdash b$ **do**
      if $occu_\alpha \leftarrow matchi(s, b.s_\alpha)$ **then**
        if not $matchf(s, b.s_\beta, occu_\alpha)$ **and not** $matchf(s, b.s_\gamma, occu_\alpha)$ **then**
          $\mathcal{D}_{u_\alpha} \leftarrow \mathcal{D}_{u_\alpha} \cup s$; $\mathcal{D}^a_{u_\alpha} \leftarrow \mathcal{D}^a_{u_\alpha} \cup subseq(s, s.begin, occu_\alpha.begin)$
          $\mathcal{D}^c_{u_\alpha} \leftarrow \mathcal{D}^c_{u_\alpha} \cup subseq(s, occu_\alpha.end, s.end)$
        end if
      end if
    end for
    /* β-inattendus */
    for all $u_\beta \vdash b$ **do**
      if $occu_\alpha \leftarrow matchi(s, b.s_\alpha)$ **and** $occu_\beta \leftarrow matchf(s, b.s_\beta, occu_\alpha, b.\tau)$ **then**
        $\mathcal{D}_{u_\beta} \leftarrow \mathcal{D}_{u_\beta} \cup s$; $\mathcal{D}^a_{u_\beta} \leftarrow \mathcal{D}^a_{u_\beta} \cup subseq(s, s.begin, occu_\alpha.begin)$
        $\mathcal{D}^c_{u_\beta} \leftarrow \mathcal{D}^c_{u_\beta} \cup subseq(s, occu_\beta.end, s.end)$
      end if
    end for
    /* γ-inattendus */
    for all $u_\gamma \vdash b$ **do**
      if $occu_\alpha \leftarrow matchi(s, b.s_\alpha)$ **and** $occu_\gamma \leftarrow matchf(s, b.s_\gamma, occu_\alpha, b.\tau)$ **then**
        $\mathcal{D}_{u_\gamma} \leftarrow \mathcal{D}_{u_\gamma} \cup s$; $\mathcal{D}^a_{u_\gamma} \leftarrow \mathcal{D}^a_{u_\gamma} \cup subseq(s, s.begin, occu_\alpha.begin)$
        $\mathcal{D}^c_{u_\gamma} \leftarrow \mathcal{D}^c_{u_\gamma} \cup subseq(s, occu_\gamma.end, s.end)$
      end if
    end for
  end for
end for
output $\mathcal{D}_u, \mathcal{D}^a_u, \mathcal{D}^c_u$ **for each** $u \vdash b \in \mathcal{B}$

expérimentations ont été réalisées sur des jeux de données réelles (Web Log) et synthétiques pour extraire des règles inattendues et étudier le passage à l'échelle des algorithmes. Le lecteur intéressé peut se reporter à Li et al. (2007).

3 Conclusion

Dans cet article, nous avons introduit la problématique de la recherche de motifs séquentiels et règles séquentielles inattendus qui trouve de très nombreuses applications dans les bases de données réelles (détection de pannes, de fraudes, de niches commerciales, etc.) L'approche USER est proposée, décomposée en différentes étapes successives (USE/USR).

Algorithm 2 Algorithme USR

Input: Base de séquences $\mathcal{D}$, base de croyances $\mathcal{B}$, ensembles de séquences produits par USE, valeurs de support minimum σ_u, σ_a, σ_c, et valeurs de confiance minimale δ_a, δ_c

Output: Ensemble P^u de motifs séquentiels inattendus, ensemble R_u^a de règles antécédentes inattendues et R_u^c de règles conséquentes inattendues pour chaque type de violation u

1: **for all** $b \in \mathcal{B}$ **do**
2: **for all** $u \vdash b$ **do**
3: $\mathcal{P}_u \leftarrow FindSequentialPatterns(\mathcal{D}_u, \sigma_u)$
4: **return** $\mathcal{P}_u$
5: $\mathcal{P}_u^a \leftarrow FindSequentialPatterns(\mathcal{D}_u^a, \sigma_a)$
6: $\mathcal{P}_u^c \leftarrow FindSequentialPatterns(\mathcal{D}_u^c, \sigma_c)$
7: **for all** $s_a \in \mathcal{P}_u^a$ **do**
8: **if** $|s_a| \,/\, |\mathcal{D}| \geq \delta_a$ **then**
9: $\mathcal{R}_u^a \leftarrow \mathcal{R}_u^a \cup \{s_a \Rightarrow u\}$
10: **end if**
11: **end for**
12: **return** $\mathcal{R}_u^a$
13: **for all** $s_c \in \mathcal{P}_u^c$ **do**
14: **if** $|s_c| \,/\, |\mathcal{D}_u| \geq \delta_a$ **then**
15: $\mathcal{R}_u^c \leftarrow \mathcal{R}_u^c \cup \{u \Rightarrow s_c\}$
16: **end if**
17: **end for**
18: **output** $\mathcal{R}_u^c$
19: **end for**
20: **end for**

Références

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

Li, D. H., A. Laurent, et P. Poncelet (2007). Découverte de motifs séquentiels et de règles inattendus. In *Internal Research Report, LIRMM*.

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*.

Summary

When considering domain knowledge within the data mining process, the most interesting sequences might not be the sequences corresponding to existing knowledge. In this paper we introduce the problem of finding unexpected behaviors within the context of sequence mining.