Fouille de motifs temporels négatifs

Katerina Tsesmeli, Manel Boumghar, Thomas Guyet, René Quiniou, Laurent Pierre

Fouille de motifs temporels négatifs

Katerina Tsesmeli*, Manel Boumghar***, Thomas Guyet**
René Quiniou*, Laurent Pierre***

*Univ Rennes, Inria, CNRS, IRISA
prenom.nom@inria.fr
**Agrocampus Ouest, IRISA - UMR 6074
***EDF R&D Saclay

Résumé. Nous étudions le problème de l'extraction de motifs fréquents contenant des événements positifs, des événements négatifs spécifiant l'absence d'événement ainsi que des informations temporelles sur le délai entre ces événements. Nous définissons la sémantique de tels motifs et proposons la méthode NTGSP basée sur des approches de l'état de l'art. Les performances de la méthode sont évaluées sur des données commerciales fournies par EDF (Électricité de France).

1 Introduction

Dans de nombreux domaines d'application tel que le diagnostic, la santé ou le marketing, les praticiens s'intéressent aux événements qui sont corrélés à des événements indésirables ou qui les déclenchent. Souvent, l'occurrence de la situation indésirable s'explique par la présence d'une action spécifique, mais aussi par l'absence de certains événements (Cao et al., 2016). Par exemple, dans le cadre du marketing, si un client de supermarché n'a pas reçu de promotions depuis longtemps, il a une très forte probabilité de choisir la concurrence, alors que dans le cas contraire il resterait fidèle à son enseigne. Pour réagir au mieux, il est important de découvrir les événements ainsi que leurs contextes d'occurrence ou d'absence afin de déterminer la meilleure action à exécuter pour éviter la situation indésirable, comme l'attrition en marketing. De plus, il est important de connaître les caractéristiques temporelles des situations indésirables, *i.e.* de quelle manière il faut anticiper l'occurrence de ces situations. Par exemple, ne pas envoyer d'offre promotionnelle dans les trois jours précédant le jour habituel de courses d'un client peut conduire ce client à acheter moins de produits, mais lui envoyer une telle offre trop tôt peut n'avoir aucun effet. Une information temporelle, comme les délais acceptables, peut améliorer la précision des prédictions et de la recommandation d'action auprès des décisionnaires.

Notre but est d'extraire d'une base de séquence d'événements datés des motifs temporels indiquant l'absence de certains événements appelés *événements négatifs* (Cao et al., 2016). Ainsi, le motif négatif $p = \langle a \neg c\, b \rangle$ indique que a est suivi fréquemment de b sans la présence de c entre eux. De plus, nous souhaitons traiter la dimension temporelle, *i.e.* savoir quel laps de temps s'écoule entre une occurrence de a et une occurrence de b en l'absence de c.

Peu de travaux se sont intéressés à la fouille de motifs séquentiels négatifs et, à notre connaissance, le présent travail est le premier qui concerne la fouille de motifs temporels négatifs. PNSP (Positive and Negative Sequential Patterns mining) (Hsueh et al., 2008), étend

GSP pour la fouille de motifs séquentiels négatifs. Toutefois, PNSP est incomplet. Sa stratégie d'élagage est incorrecte et nombre de motifs séquentiels négatifs pourtant fréquents ne sont pas extraits. Neg-GSP (Zheng et al., 2009) propose une version alternative s'appuyant sur une application partielle du principe d'Apriori aux motifs séquentiels positifs mais pas aux négatifs. Cependant la stratégie d'élagage de Neg-GSP bien que correcte est très inefficace car elle n'applique le principe d'Apriori qu'à la partie positive du motif. Proposé récemment, e-NSP (efficient NSP, Cao et al. (2016)) calcule le support des motifs séquentiels négatifs à partir du support de leurs sous-motifs séquentiels positifs, ce qui ne nécessite pas de parcours supplémentaire de la base de séquences. e-NSP produit les mêmes motifs que PNSP et souffre de la même incomplétude mais il est nettement plus performant que PNSP et Neg-GSP.

Nous proposons une formalisation et une sémantique pour les motifs temporels négatifs. Nous proposons une méthode pour résoudre le problème d'extraction de tels motifs à partir d'une base de séquences temporelles. La méthode d'extraction proposée s'appuie sur les algorithmes PrefixSpan (Pei et al., 2004), e-NSP (Cao et al., 2016) et TGSP (Yen et Lee, 2013). Il faut noter qu'aucun de ces algorithmes ne traite complètement le problème d'extraction de motifs qui nous intéresse. Enfin, la méthode est évaluée sur des données réelles.

2 Motifs temporels négatifs

Cette section introduit les motifs temporels négatifs qui étendent les motifs séquentiels (Pei et al., 2004), d'une part, avec des contraintes imposant l'absence de certains itemsets (aspect négatif) et, d'autre part, des contraintes sur le délai entre occurrences d'itemsets (aspect temporel). Nous proposons un formalisme pour de tels motifs et définissons leur sémantique. Dans la suite, $[n] = \{1, \ldots, n\}$ dénote l'ensemble des n premiers entiers strictement positifs.

Soit $\mathcal{I}$ un ensemble d'items. Formellement, un *motif temporel négatif (MTN)* est une séquence $\boldsymbol{p} = \langle \xrightarrow[\neg q_0]{[l_0,u_0]} p_1 \xrightarrow[\neg q_1]{[l_1,u_1]} p_2 \xrightarrow[\neg q_2]{[l_2,u_2]} \ldots \xrightarrow[\neg q_{n-1}]{[l_{n-1},u_{n-1}]} p_n \xrightarrow[\neg q_n]{[l_n,u_n]} \rangle$ où $p_{i_{i\in[n]}}$ est un itemset positif ($p_i = \{p_i^j,\ p_i^j \in \mathcal{I}$), $\neg q_{i_{i\in[0,n]}}$ est un itemset négatif ($q_i = \{q_i^j,\ q_i^j \in \mathcal{I}$) et $[l_i, u_i]_{i\in[0,n]}, l_i, u_i \in \mathbb{R}^+$ est une contrainte temporelle spécifiant les bornes du délai admissible entre l'occurrence de p_i et celle de p_{i+1}. $[l_0, u_0]$ (resp. $[l_n, u_n]$) encadre le délai entre le début de la séquence et l'occurrence de p_1 (resp. entre l'occurrence de p_n et la fin de la séquence). Une contrainte d'absence (itemset négatif) peut être vide, de même qu'une contrainte temporelle (intervalle temporel).

À noter qu'un itemset, positif ou négatif, ne contient que des items positifs. De plus, un motif ne peut contenir deux itemsets négatifs consécutifs. En effet, une telle succession est difficilement interprétable (cf. Cao et al. (2016)). La *partie positive* du motif est la sous-séquence restreinte aux itemsets positifs.

Le motif temporel négatif $\boldsymbol{p} = \langle \xrightarrow[\neg q_0]{[l_0,u_0]} p_1 \xrightarrow[\neg q_1]{[l_1,u_1]} \ldots \xrightarrow[\neg q_{n-1}]{[l_{n-1},u_{n-1}]} p_n \xrightarrow[\neg q_n]{[l_n,u_n]} \rangle$ a une *occurrence* dans la séquence $\boldsymbol{s} = \langle s_1, \ldots, s_m \rangle$ (s *supporte* $\boldsymbol{p}$) ssi **il existe** des indices $(e_i)_{i\in[n]}$ tels que $\forall i \in [n],\ e_i \in [m] \wedge j < k \Rightarrow e_j < e_k$ et :

1. $\forall i \in [1, n],\ p_i \subseteq s_{e_i}$ (inclusion des itemsets positifs),
2. $\forall i \in [0, n],\ \forall j, e_i < j < e_{i+1},\ q_i \nsubseteq s_j$ (satisfaction des contraintes d'absence). e_0 (resp. e_{n+1}) est un indice virtuel marquant le début (resp. la fin) de la séquence $\boldsymbol{s}$.
3. $\forall i \in [0, n],\ l_{e_i} \leq t_{e_{i+1}} - t_{e_i} \leq u_{e_i}$ (satisfaction des contraintes temporelles),

Soient $\neg q_i$ un itemset négatif de $\boldsymbol{p}$, p_i son itemset (positif) prédécesseur (ou un marqueur de début de séquence) et p_{i+1} son itemset (positif) successeur (ou un marqueur de fin de séquence). La définition précédente spécifie l'*absence faible* pour les itemsets négatifs : **il existe une** occurrence de (p_i, p_{i+1}) dans $\boldsymbol{s}$ qui ne contient pas q_i. L'absence forte impose que **toute** occurrence de (p_i, p_{i+1}) dans $\boldsymbol{s}$ ne contient pas q_i.

Exemple 1 (Occurrence d'un motif séquentiel négatif (MSN) sous sémantique d'absence forte et faible)**.** Soient le motif $\boldsymbol{p} = \langle a \rightarrow b \underset{\neg c}{\longrightarrow} d\rangle$ et les séquences $\boldsymbol{s_1} = \langle a\ b\ e\ d\rangle$ et $\boldsymbol{s_2} = \langle a\ b\ c\ a\ d\ e\ b\ d\rangle$. La partie positive de $\boldsymbol{p}$ est $\langle a\ b\ d\rangle$. Elle a une seule occurrence dans $\boldsymbol{s_1}$, par conséquent les deux sémantiques ne font pas de différence. Mais elle a 4 occurrences dans $\boldsymbol{s_2}$ aux positions $(1, 2, 5)$, $(1, 2, 8)$, $(1, 7, 8)$ et $(4, 7, 8)$. Les deux premières occurrences ne satisfont pas la contrainte négative ($\neg c \equiv$ absence de c entre b et d) alors que les deux dernières la satisfont. Selon la sémantique d'absence faible, la séquence $\boldsymbol{s_2}$ contient le motif p alors que selon la sémantique d'absence forte la séquence $\boldsymbol{s_2}$ ne contient pas le motif p.

Exemple 2 (Occurrence d'un motif temporel négatif)**.** Le MTN $\boldsymbol{p} = \langle a \underset{\neg c}{\overset{[1,3]}{\longrightarrow}} e \overset{[2,2]}{\longrightarrow} d\rangle$ spécifie qu'il y a entre 1 et 3 unités de temps entre l'occurrence de l'item a et l'occurrence de l'item e et il ne doit pas y avoir d'occurrence de l'item c entre l'occurrence de a et celle de e. De plus, l'item d doit se produire exactement 2 unités de temps après e.

La séquence temporelle $\boldsymbol{s} = \langle (d, 37), (a, 38), (e, 41), (b\ c, 42), (d, 43)\rangle$ supporte $\boldsymbol{p}$ (ou $\boldsymbol{p}$ a une occurrence dans $\boldsymbol{s}$). Les itemsets positifs de $\boldsymbol{p}$ apparaissent aux positions $(2, 3, 5)$ sans occurrence de c entre a (position 2) et e (position 3). L'occurrence de e, 3 unité de temps après a, satisfait la contrainte temporelle entre a et e. De même, l'occurrence de d, 2 unité de temps après l'occurrence de e, satisfait la contrainte temporelle entre e et d.

La fouille de motifs temporels négatifs dans une base de séquences temporelles datées $\mathcal{D}$ consiste à extraire toutes les sous-séquences (motifs) incluses fréquemment dans des séquences de la base, *i.e.* ayant un support supérieur à un seuil σ donné a priori. Pour réduire la complexité de la recherche, nous reprenons la contrainte d'e-NSP imposant que tout itemset négatif doit être fréquent. Contrairement à e-NSP, la négation spécifie l'absence faible.

3 NTGSP : fouille de motifs temporels négatifs

Dans cette section nous présentons la méthode NTGSP – Negative Time Gap Sequential Pattern – pour l'extraction des motifs temporels négatifs (MTN) à partir de séquences d'itemsets datés. NTGSP emprunte à PrefixSpan (Pei et al., 2004) pour l'extraction de motifs séquentiels, à e-NSP (Cao et al., 2016) pour l'extraction de motifs négatifs et à TGSP (Yen et Lee, 2013) pour l'extraction d'intervalles temporels entre itemsets positifs.

Les quatre étapes de l'algorithme sont décrites ci-dessous. Le processus de fouille sera illustré sur la base $\mathcal{D}$ suivante et les paramètres $\sigma = 2$, $\mu = 3$, $\varepsilon = 1$ et $\delta = 2$:

$\boldsymbol{s}_1$:	$\langle((ab), 1), (a, 4), (d, 7), (d, 9)\rangle$	$\boldsymbol{s}_2$:	$\langle((ab), 2), (b, 3), (d, 9)\rangle$
$\boldsymbol{s}_3$:	$\langle((ab), 4), (c, 5), (a, 7), (d, 10)\rangle$	$\boldsymbol{s}_4$:	$\langle(b, 3), ((ab), 5), (d, 11)\rangle$
$\boldsymbol{s}_5$:	$\langle((ab), 3), (c, 7), (d, 9)\rangle$		

1. Extraction des motifs séquentiels La première étape utilise l'algorithme PrefixSpan (Pei et al., 2004) avec le seuil de support minimum σ pour extraire les motifs séquentiels de la base de séquences temporelles $\mathcal{D}$.

Les motifs séquentiels de $\mathcal{D}$ sont $\langle (ab)\ c\ d \rangle$, $\langle b\ b\ d \rangle$ et tous leurs sous-motifs.

2. Génération des motifs séquentiels négatifs candidats (MSC) Les MSC sont générés en passant à négatif un ou plusieurs itemsets de chacun des motifs générés à l'étape précédente.

Les MSC générés à partir du motif $\langle (ab)\ c\ d \rangle$ de l'étape 1 sont :
$\boldsymbol{p_1} = \langle \xrightarrow[\neg(ab)]{} c \rightarrow d \rangle$, $\boldsymbol{p_2} = \langle (ab) \underset{\neg c}{\longrightarrow} d \rangle$, $\boldsymbol{p_3} = \langle (ab) \rightarrow c \underset{\neg d}{\longrightarrow} \rangle$, $\boldsymbol{p_4} = \langle \xrightarrow[\neg(ab)]{} c \underset{\neg d}{\longrightarrow} \rangle$.

3. Extraction des motifs séquentiels négatifs (MSN) NTGSP calcule le support sur le dataset $\mathcal{D}$ de chaque MSC obtenu à l'étape 2 et élague les MSC dont le support est inférieur au seuil μ fixant le le support minimum des MSN.

Parmi les MSC de l'étape 2, seul $\boldsymbol{p_2} = \langle (ab) \underset{\neg c}{\longrightarrow} d \rangle$ a un support strictement positif : $supp(\boldsymbol{p_2}) = |\{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_4\}| = 3 \geq \mu$. $\boldsymbol{p_2}$ est donc un motif séquentiel négatif.

4. Extraction des motifs temporels négatifs (MTN) Finalement, NTGSP extrait les intervalles temporels des délais admissibles entre les itemsets positifs des MSN par clustering des délais fournis par les occurrences de motifs dans les séquences temporelles de $\mathcal{D}$. Pour ce faire NTGSP utilise l'algorithme de clustering CLIQUE (Agrawal et al., 2005). Soit $\boldsymbol{p}$ un MSN obtenu à l'étape 3 et comportant n itemsets positifs. Pour chaque occurrence de $\boldsymbol{p}$ dans une séquence de $\mathcal{D}$, les délais entre itemsets positifs de $\boldsymbol{p}$ peut se représenter par un point dans un espace de dimension $n-1$. Nous cherchons à regrouper ces points proches pour construire la composante temporelle des MTN. CLIQUE décompose l'espace de dimension $n-1$ en hypercubes unitaires de taille ε et élague ceux qui ne sont pas suffisamment denses relativement à un seuil δ. Chaque composante connexe du graphe des hypercubes unitaires denses constitue un cluster. Un cluster est ensuite décomposé en sous-hypercubes maximaux et les contraintes temporelles sont générées à partir des coordonnées des côtés de ces sous-hypercubes dans chaque dimension. Pour des raisons d'efficacité NTGSP utilise la sémantique d'*absence faible* et seule la première occurrence du motif dans une séquence temporelle est utilisée.

Soit $\varepsilon = 1$ et $\delta = 2$. Le vecteur de délais associé à $\boldsymbol{p_3} = \langle (ab) \underset{\neg c}{\longrightarrow} d \rangle$ est de dimension 1. Aux occurrences de $\boldsymbol{p_3}$ sont associés respectivement les vecteurs de délais $(6), (7)$ et (6). Seule l'unité 6 est dense : elle contient 2 points. Elle constitue un cluster. La contrainte temporelle résultante est $[6, 6]$. Nous obtenons donc le motif $\langle (ab) \overset{[6,6]}{\underset{\neg c}{\longrightarrow}} d \rangle$.

4 Expérimentations

NTGSP a été expérimenté sur des données synthétiques et sur des données réelles fournies par EDF. Faute de place nous présentons des résultats sur des données réelles uniquement.

NTGSP est implémenté en C++ (algorithme PrefixSpan) et Python (NSP et TGSP). L'identification des occurrences de motifs séquentiels négatifs est effectuée par un algorithme de recherche d'expressions régulières car ces algorithmes bénéficient d'années de recherche et sont très efficaces.

Le jeu de données CRM (Customer Relationship Management) d'EDF contient les séquences temporelles correspondant aux interactions de 375.143 clients avec la compagnie, soit globalement 944.514 itemsets datés (sur 75 items). La figure 1 illustre quelques aspects quan-

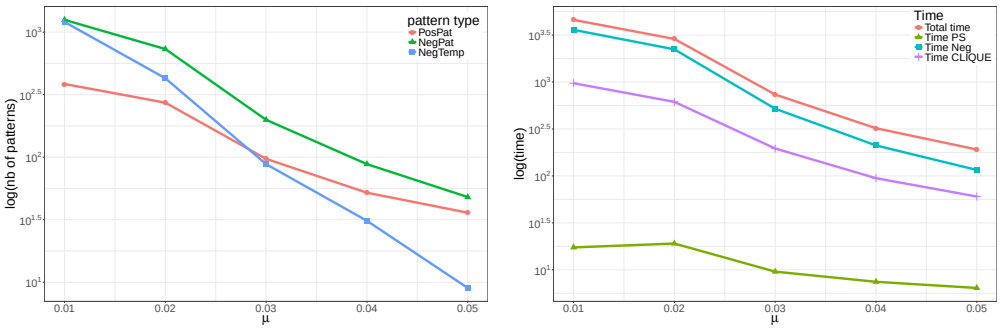

FIG. 1 – *Nombre de motifs positifs, négatifs et temporels, à gauche, et temps d'exécution (en secondes), à droite, en fonction du support minimum (μ).*

titatifs de la fouille de motifs montrant que le nombre de motifs et temps d'exécution suivent une classique complexité exponentielle lorsque le support diminue.

Le tableau 1 fournit quelques exemples de motifs extraits des données EDF. Pour des raisons de confidentialité les événements ont dû être anonymisés. Cependant, pour montrer l'intérêt de NTGSP, le tableau contient une classe de motifs particulièrement importante, que nous avons appelés *motifs duaux*. Ils se présentent par paires $(\langle C\ \neg A\ U\rangle,\ \langle C\ A\ \neg U\rangle)$ où C dénote un contexte, A un événement lié à une action ou un état et U un *événement indésirable*, lié à une panne, par exemple. Le premier motif de la paire exprime que dans le contexte C et en l'absence de l'action A l'événement indésirable U se produit. Le deuxième motif indique dans le même contexte en présence de A l'événement indésirable U ne se produit pas.

Les motifs 1 et 2 sont de tels motifs duaux. De plus, ils fournissent des informations temporelles importantes pour savoir dans quel laps de temps l'action inhibitrice doit être exécutée pour éviter l'événement indésirable. À noter le support élevé des motifs 1 et 2 : le motif 1 se produit dans $0.0562 \times 375.143 = 21.083$ interactions de clients et le motif 3 a un support de 4.0% et apparaît dans 15.000 séquences, ce qui est particulièrement significatif.

Les motifs (non duaux) 3 et 4 contiennent la même séquence d'événements mais des contraintes temporelles différentes qui correspondent à deux hypercubes (rectangles) maximaux issus de deux représentations différentes d'un cluster.

TAB. 1 – *Quelques exemples motifs temporels négatifs*

num	motif négatif	motif temporel négatif	support
1	$\langle 4\ \neg 3\ 10\rangle$	$4 \xrightarrow[\neg 3]{[0,190]} 10$	5.62%
2	$\langle 4\ 3\ \neg 10\rangle$	$4 \xrightarrow{[1,240]} 3 \xrightarrow[\neg 10]{}$	4.00%
3	$\langle 10\ 3\ 10\ \neg 4\rangle$	$10 \xrightarrow{[1,50]} 3 \xrightarrow{[0,30]} 10 \xrightarrow[\neg 4]{}$	1.03%
4	$\langle 10\ 3\ 10\ \neg 4\rangle$	$10 \xrightarrow{[1,60]} 3 \xrightarrow{[0,20]} 10 \xrightarrow[\neg 4]{}$	1.06%

5 Conclusion et perspectives

Dans cet article nous avons présenté l'algorithme NTGSP pour résoudre le problème de l'extraction de motifs temporels négatifs à partir d'une base de séquences temporelles datées. NTGSP combine différentes méthodes de l'état de l'art, mais c'est la première fois que de telles méthodes sont associées pour résoudre un problème qui s'avère important d'un point de vue applicatif. L'autre contribution importante de ce travail est la formulation d'une sémantique pour les motifs temporels négatifs, mixant négation et contraintes temporelles.

La solution proposée s'est avérée adéquate pour évaluer l'intérêt des motifs temporels négatifs sur une base de séquence réelle, mais deux points doivent être améliorés. Le premier concerne la structure de l'algorithme qui effectue les opérations en séquence au lieu de « pousser » les contraintes d'absence et les contraintes temporelles dans le processus de fouille de motifs séquentiels initial. Le second concerne la complétude de la proposition. Nous souhaitons, à l'avenir, proposer une alternative aux contraintes de la méthode d'e-NSP, notoirement incomplète pour la génération de motifs séquentiels négatifs.

Références

Agrawal, R., J. Gehrke, D. Gunopulos, et P. Raghavan (2005). Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery 11*(1), 5–33.

Cao, L., X. Dong, et Z. Zheng (2016). e-NSP : Efficient negative sequential pattern mining. *Artificial Intelligence 235*, 156–182.

Hsueh, S.-C., M.-Y. Lin, et C.-L. Chen (2008). Mining negative sequential patterns for e-commerce recommendations. In *Proceedings of Asia-Pacific Services Computing Conference*, pp. 1213–1218. IEEE.

Pei, J., J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, et M.-C. Hsu (2004). Mining Sequential Patterns by Pattern-Growth : The PrefixSpan Approach. *IEEE Transactions on knowledge and data engineering 16*(11), 1424–1440.

Yen, S.-J. et Y.-S. Lee (2013). Mining non-redundant time-gap sequential patterns. *Applied Intelligence 39*(4), 727–738.

Zheng, Z., Y. Zhao, Z. Zuo, et L. Cao (2009). Negative-GSP : An efficient method for mining negative sequential patterns. In *Proceedings of the Australasian Data Mining Conference*, pp. 63–67.

Summary

We investigate the problem of extracting frequent sequences with positive events, negative events specifying the absence of events as well as temporal information about the delay between these events. We formulate the semantics of such patterns and we propose algorithm NTGSP based on state of the art methods. The performance of NTGSP is evaluated on commercial data provided by EDF, a major french power distribution company.