Analyse d'une enquête sur la sémantique des motifs séquentiels avec négation

Thomas Guyet

Analyse d'une enquête sur la sémantique des motifs séquentiels avec négation

Thomas Guyet[1]

[1] Inria – Centre de Lyon, AIstroSight

thomas.guyet@inria.fr

Résumé

Un motif séquentiel avec négation prend la forme d'un motif séquentiel pour lequel le symbole de négation peut être utilisé devant certains des itemsets. Dans ce cas, l'itemset qui suit doit être absent d'une séquence pour que le motif apparaisse dans cette séquence. Des travaux récents ont montré que différentes sémantiques pouvaient être attribuées à ces formes de motif. Ces travaux ont ainsi mis en évidence que les algorithmes d'extraction de ces motifs n'extrayaient pas les mêmes ensembles de motifs et ils soulèvent la question de l'interprétabilité des résultats. Dans ce travail, nous nous sommes demandés si certaines sémantiques étaient plus intuitives que d'autres et si celles-ci correspondaient à celles d'un ou plusieurs algorithmes de l'état de l'art. Pour cela, nous avons procédé sous la forme d'un questionnaire. Cet article présente ce questionnaire et l'analyse des 124 réponses. Les résultats montrent que deux sémantiques sont majoritaires mais qu'aucune d'elles ne correspond à celles des algorithmes principaux de l'état de l'art. Des recommandations sont faites pour tenir compte de ce résultat.

Mots-clés

Extraction de motifs, motifs séquentiels, négation, interprétation, enquête.

Abstract

A sequential pattern with negation takes the form of a sequential pattern for which the negation symbol can be used before some of the itemsets. In this case, the following itemset must be absent in a sequence for the pattern to appear in this sequence. Recent work has shown that these patterns have different semantics and raises the question of the interpretability of pattern mining algorithms. This article presents a questionnaire about the intuitiveness of some semantics. The analysis of the 124 answers shows that there are mainly two semantics that are mostly intuitive but that none of them corresponds to those of the main algorithms of the state of the art. Recommendations are made to address this outcome.

Keywords

Pattern mining, sequential patterns, negation, survey.

1 Introduction

L'extraction de motifs séquentiels est une classe de méthodes classiques de la fouille de données. Elle vise à extraire des sous-séquences (motifs) qui apparaissent fréquemment dans une grande base de séquences. Le motif apparaît fréquemment s'il apparaît dans au moins σ séquences, où σ est défini par l'utilisateur. Par exemple, prenons le motif $\langle e\ (ca)\ d\rangle$ désignant que "l'item e est suivi de a et c en même temps (*itemset*) puis de d"[1]. Dans le tableau ci-dessous, ce motif apparaît dans 4 séquences ($\boldsymbol{p_0}$, $\boldsymbol{p_2}$, $\boldsymbol{p_3}$ et $\boldsymbol{p_4}$).

Table 1: Exemple de base de séquences. La case à cocher sur la droite permet au lecteur de répondre lui-même aux questions. Voir Question 1 pour ce tableau.

id	*Séquence*	
$\boldsymbol{p_0}$	$\langle e\ (caf)\ d\ b\ e\ d\rangle$	☐
$\boldsymbol{p_1}$	$\langle c\ a\ d\ b\ e\ d\rangle$	☐
$\boldsymbol{p_2}$	$\langle e\ (ca)\ d\rangle$	☐
$\boldsymbol{p_3}$	$\langle d\ e\ (ca)\ b\ d\ b\ e\ f\rangle$	☐
$\boldsymbol{p_4}$	$\langle c\ e\ b\ (fac)\ d\ e\ c\rangle$	☐

Ces motifs fréquents peuvent être énumérés efficacement, grâce à la propriété d'anti-monotonie du support (*i.e.* le nombre d'occurrences d'un motif). Intuitivement, le support d'un motif décroît avec la taille des motifs. Cette propriété, utilisée par la plupart des algorithmes de la littérature, évite d'énumérer les motifs qui sont plus grands que des motifs qu'on sait a priori ne pas être fréquents.

Plusieurs travaux [4, 7] ont enrichi le domaine des motifs séquentiels par l'ajout d'information sur l'absence de la survenue d'un évènement. On parle alors de motifs séquentiels *avec négation*. Les motifs séquentiels avec négation prennent la forme de motifs séquentiels pour lesquels un symbole de négation, $\neg$, devant un itemset indique que ce dernier doit être absent d'une séquence pour y apparaître. Intuitivement, le motif $\langle a\ \neg b\ c\rangle$ sera reconnu dans une séquence si cette dernière comporte un a puis un c et que b est absent entre les occurrences de a et c.

Néanmoins, il a été constaté que les deux algorithmes principaux eNSP [4] et NegPSpan [7] n'extraient pas les

[1] On suppose ici que (ca) et (ac) désignent le même ensemble d'items.

mêmes ensembles de motifs négatifs. Ceci s'explique par le fait que ces deux algorithmes n'attribuent pas la même sémantique au symbole de négation [2].[2] Pour un motif $\boldsymbol{p}$ et une séquence $\boldsymbol{s}$ donnés, eNSP et NEGPSPAN ne seront pas forcément d'accord sur le fait que $\boldsymbol{p}$ apparaisse ou non dans $\boldsymbol{s}$. Les comptages d'apparition dans la base sont donc différents pour ces algorithmes et les motifs qu'ils considèrent comme effectivement fréquents peuvent ainsi être différents.

Les deux sémantiques sont toutes aussi intéressantes l'une que l'autre. La question qui se pose réside alors sur le partage de la sémantique entre l'utilisateur et l'outil qu'il utilise. Autrement dit : l'utilisateur à qui sont délivrés les motifs a-t-il une interprétation similaire à celle de l'algorithme utilisé ? Si ce n'est pas le cas, il peut y avoir une mauvaise interprétation des résultats de l'extraction de motifs. Un utilisateur non-expert ne cherchant pas forcément à comprendre les subtilités de ces motifs, il semble utile d'identifier une possible disparité entre la sémantique utilisée dans un algorithme et celle utilisée "intuitivement" par un utilisateur. Si cette disparité existe, il sera alors nécessaire de proposer des solutions évitant une mauvaise interprétation des résultats.

Dans cet article, nous nous sommes donc principalement posé trois questions :

1. existe-t-il une sémantique "intuitive" pour les motifs avec négation ?
2. la sémantique "intuitive" correspond-elle à celle qui est effectivement utilisée par l'un des algorithmes de l'état de l'art ?
3. quelles recommandations faire sur l'usage des motifs avec négation ?

Pour répondre à ces questions, la méthodologie a consisté à proposer un questionnaire pour révéler la sémantique qui est intuitivement appliquée par les utilisateurs. Le détail de la méthodologie de cette enquête est décrit dans la Section 3. La Section 5 présente les questions qui ont été posées aux utilisateurs et explicite les interprétations alternatives qui sont possibles. La Section 6 présente et analyse les résultats qui ont été collectés auprès de 124 participants. Avant cela, on commence par un bref état de l'art des méthodes d'extraction de motifs séquentiels avec négation.

2 État de l'art sur l'extraction de motifs séquentiels avec négation

Les premiers travaux sur l'extraction de motifs négatifs ont été proposés par Savasere et *al.* [11] dans le cadre de la fouille d'itemsets. Les premiers travaux sur les motifs séquentiels avec négation ont été proposés par Wu et *al.* [14] pour des règles d'association. Plusieurs approches récentes ont été proposées pour bénéficier également des avancées dans le domaine de l'extraction de motifs. L'algorithme eNSP extrait des motifs négatifs en exploitant des opérations ensemblistes entre motifs séquentiels fréquents [4]. Il évite ainsi l'énumération directe des motifs avec négation, car l'ensemble des motifs qui sont extraits ne bénéficient pas de la propriété d'antimonotonie. De nombreuses variantes de cet algorithme ont été proposées depuis, s'intéressant à l'utilité des items [15], aux répétitions [5], aux contraintes multiples de supports [16], etc. NEGPSPAN [7] est une approche concurrente à eNSP qui utilise une sémantique de motifs différente. Cette sémantique bénéficie de la propriété d'antimonotonie. Ceci permet une extraction efficace et complète selon les principes classiques de l'extraction de motifs. Récemment, Wang et *al.* [13] ont proposé VM-NSP, un algorithme qui utilise une représentation verticale pour améliorer l'efficacité des algorithmes. Le lecteur intéressé par un état de l'art plus complet des approches récentes d'extraction de motifs séquentiels avec négation peut se référer à Wang et *al.* [12].

Les premières approches se sont comparées entre elles bien qu'elles n'utilisent pas les mêmes sémantiques de motifs. L'identification des différentes sémantiques a conduit à clarifier le domaine [2]. Plus précisément, huit sémantiques des motifs avec négation ont été identifiées. Ces huit sémantiques sont issues de choix possibles d'interprétation de la notion de non-inclusion, d'occurrence et de relation d'inclusion. La section 5 détaille ces notions.

3 Enquête sur la perception des motifs avec négation

L'enquête[3] mise en place vise à identifier une sémantique qui serait plus naturellement utilisée par les utilisateurs d'algorithmes d'extraction de motifs. Cette enquête est organisée en trois parties (la section suivante revient plus en détail sur les questions des phases 2 et 3 de l'enquête) :

1. estimation du niveau de connaissance du domaine de la fouille de motifs et de la logique. Dans cette partie, on demande si l'utilisateur est familier des notions d'extraction de motifs, et également s'il est informaticien/logicien/chercheur. L'objectif de cette question est de disposer d'informations pour caractériser d'éventuels biais de l'ensemble des enquêtés.
2. vérification de la compréhension des principes des motifs séquentiels afin de limiter les biais de compréhension dans la suite des questions. Tout d'abord, un texte explique et illustre les principes des motifs séquentiels. Une première question évalue la compréhension de la sémantique des motifs séquentiels (sans négation), notamment les notions d'*itemset*, le séquencement et la possibilité de *gaps*[4].

[2] Ce n'est pas la seule raison de la divergence entre les algorithmes. Mais les autres différences sont mineures.

[3] Enquête : `http://people.irisa.fr/Thomas.Guyet/negativepatterns/Survey/survey.php`

[4] La reconnaissance de la sous-séquence permet l'insertion d'itemsets au milieu d'une occurrence.

Tant que la réponse à cette question n'est pas correcte, l'utilisateur ne peut pas poursuivre le questionnaire. Une seconde question vérifie que la portée des négations est comprise telle que définie par notre cadre d'analyse [2]. Par exemple, pour le motif $\langle a\ \neg b\ c\rangle$, la négation du b ne porte pas au-delà d'une occurrence de c. Ainsi, ce motif est considéré comme apparaissant dans la séquence $\langle a\ e\ c\ b\rangle$ même si un b apparaît après le c. Les utilisateurs ne répondant pas correctement à cette question seront écartés de l'analyse des réponses.

3. identification de la sémantique « intuitive » des motifs séquentiels avec négation. Pour chacune de ces questions, on demande à l'utilisateur de cocher les séquences dans lesquelles il pense qu'un motif apparaît (voir exemple de la Figure 1). Le groupe de séquences cochées associe donc un utilisateur à une sémantique donnée.

L'enquête a été diffusée au travers de listes de diffusion de recherche ainsi que dans des cercles non liés à la recherche pour collecter des réponses de non-experts. Elle est accessible via un navigateur web standard. Le questionnaire est rédigé en anglais et s'adresse donc à des anglophones. Les explications relatives aux principes de la notion de motif séquentiel sont détaillées en début de questionnaire. Afin d'éviter le biais de maîtrise des représentations mathématiques, le questionnaire peut être joué en deux versions : avec notations sous forme de lettres ou de symboles colorés (cf. Figure 1).
Le questionnaire est totalement anonyme. Seule la date de saisie du questionnaire a été collectée en sus des réponses aux questions.

4 Cadre général

On commence par introduire la syntaxe des motifs séquentiels avec négation. Dans toute la suite, $[n] = \{1, \ldots, n\}$ désigne l'ensemble des n premiers entiers, et $\mathcal{I}$ désigne un ensemble d'items (alphabet). Un sous-ensemble $A = \{a_1\ a_2\ \cdots\ a_m\} \subseteq \mathcal{I}$ est nommé un *itemset*. Une *séquence* $\boldsymbol{s}$ est de la forme $\boldsymbol{s} = \langle s_1\ s_2\ \cdots\ s_n\rangle$ où s_i est un itemset.

Définition 1 (Motif séquentiel avec négation)**.** *Un motif séquentiel avec négation* $\boldsymbol{p} = \langle p_1\ \neg q_1\ p_2\ \neg q_2\ \cdots\ p_{n-1}\ \neg q_{n-1}\ p_n\rangle$ *est telle que* $p_i \in 2^{\mathcal{I}} \setminus \{\emptyset\}$ *pour tout* $i \in [n]$ *et* $q_i \in 2^{\mathcal{I}}$ *pour tout* $i \in [n-1]$. $\boldsymbol{p}^+ = \langle p_1\ p_2\ \cdots p_n\rangle$ *désigne la partie positive de* $\boldsymbol{p}$.

La sémantique des motifs repose sur la relation d'inclusion. Cette relation précise comment considérer si un motif apparaît (est inclus) ou non dans une séquence. Cette relation utilise la notion d'occurrence d'un motif dans une séquence, formellement définie ainsi :

Définition 2 (Occurrence d'un motif séquentiel)**.** *Soit une séquence* $\boldsymbol{s} = \langle s_1 \ldots s_n\rangle$ *et* $\boldsymbol{p} = \langle p_1 \ldots p_m\rangle$ *un motif séquentiel.* $\boldsymbol{e} = (e_i)_{i\in[m]} \in [n]^m$ *est une* occurrence *du motif* $\boldsymbol{p}$ *dans la séquence* $\boldsymbol{s}$ *ssi* $\forall i \in [m],\ p_i \subseteq s_{e_i}$ *et* $e_i < e_{i+1}$ *pour tout* $i \in [m-1]$.

Partant de cette définition, nous avons construit la question suivante pour en valider sa compréhension avant de poursuivre la suite du questionnaire.

Question 1 (Occurrence d'un motif séquentiel)**.** *Soit le motif séquentiel* $\boldsymbol{p} = \langle (ca)\ d\ e\rangle$*, indiquer dans quelles séquences de la Table 1 apparaît le motif* $\boldsymbol{p}$.[5]

Les réponses attendues à cette question sont les séquences $\boldsymbol{p_0}$, $\boldsymbol{p_3}$ et éventuellement $\boldsymbol{p_4}$. La séquence $\boldsymbol{p_0}$ permet de vérifier la compréhension que (ca) apparaît dans (caf) selon nos définitions. La séquence $\boldsymbol{p_1}$ permet de vérifier qu'il faut que tous les éléments de (ca) apparaissent ensemble. La séquence $\boldsymbol{p_2}$ permet vérifier la compréhension de l'importance de l'ordre dans la séquence. La séquence $\boldsymbol{p_3}$ permet vérifier la compréhension de la notion de *gap* : il est possible d'avoir des itemsets au milieu d'une occurrence (par exemple, la survenue de b entre le d et le e). Finalement, la dernière séquence présente un itemset dont les items ne sont pas ordonnés. Dans le cas où $\boldsymbol{p_4}$ ne serait pas jugé contenir $\boldsymbol{p}$ alors on serait informé d'une sensibilité de l'utilisateur à l'ordre présenté dans un itemset (ce qui n'est classiquement pas le cas).
De la même manière, la sémantique des motifs séquentiels avec négation repose sur une relation d'inclusion. Un motif avec négation, $\boldsymbol{p}$, est inclus dans une séquence $\boldsymbol{s}$ si $\boldsymbol{s}$ contient une sous-séquence $\boldsymbol{s}'$ telle que chaque ensemble positif de $\boldsymbol{p}$, *i.e.* p_i, est inclus dans un itemset de $\boldsymbol{s}'$ (en respectant l'ordre) et que toutes les contraintes de négations exprimées par les $\neg q_i$ sont également satisfaites. La contrainte de négation de q_i s'appliquant alors à la sous-séquence de $\boldsymbol{s}'$ située entre l'occurrence de l'itemset positif précédant $\neg q_i$ dans $\boldsymbol{p}$ et l'occurrence de l'itemset positif suivant $\neg q_i$ dans $\boldsymbol{p}$.
Cette définition détermine la portée de la négation. Cette définition est propre au cadre dans lequel nous travaillons par la suite. Aussi, il est important de vérifier qu'il est partagé par les utilisateurs. La question suivante permet de s'en assurer.

Question 2 (Portée de la négation)**.** *On considère un motif* $\boldsymbol{p} = \langle c\ \neg d\ e\rangle$*. Indiquer les séquences de la base ci-dessous dans lesquelles, selon vous,* $\boldsymbol{p}$ *apparaît.*

id	*Séquence*	
$\boldsymbol{s_0}$	$\langle f\ f\ c\ b\ d\ a\ e\rangle$	☐
$\boldsymbol{s_1}$	$\langle f\ c\ b\ f\ a\ e\rangle$	☐
$\boldsymbol{s_2}$	$\langle b\ f\ c\ b\ a\rangle$	☐
$\boldsymbol{s_3}$	$\langle b\ c\ b\ e\ d\rangle$	☐
$\boldsymbol{s_4}$	$\langle f\ a\ c\ e\ b\rangle$	☐

Dans cette question, il est raisonnable de considérer que $\boldsymbol{p}$ apparaît dans $\boldsymbol{s_1}$, $\boldsymbol{s_3}$ (le d est hors de la portée supposée de la négation) et $\boldsymbol{s_4}$. Les enquêté(e)s qui ne cochent pas $\boldsymbol{s_4}$ ont probablement interprété la contrainte $\neg d$ comme : l'apparition d'un élément qui n'est pas d (ce qui n'est pas dans les définitions proposées par la suite).

[5] Le lecteur est invité à remplir lui-même les réponses aux questions dans les tableaux avant de lire les explications.

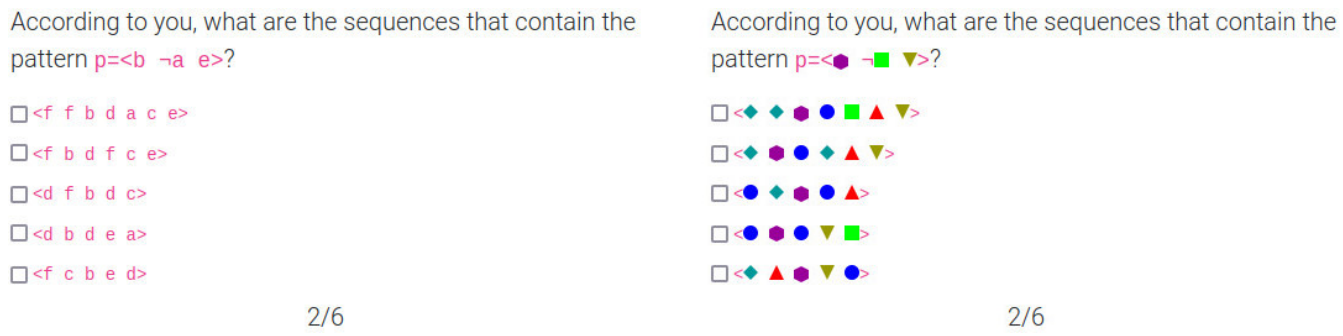

Figure 1: Illustration des deux versions du questionnaire : sur la gauche, version avec lettres, sur la droite, version avec symboles colorés. La question consiste à désigner les séquences pour lesquels l'utilisateur pense qu'elle contient un motif.

Si $\boldsymbol{p_0}$ est considéré comme contenant $\boldsymbol{p}$ il est probable que la contrainte $\neg d$ soit comprise comme devant suivre strictement après c (de nouveau, ce n'est pas une situation considérée dans l'analyse de Besnard et Guyet [2]).

5 Questions sur la sémantique des négations

Dans cette section, nous reprenons les questions de la troisième partie du questionnaire et nous expliquons les différentes interprétations que révèlent les réponses faites par les enquêtés.

5.1 Non-inclusion d'un itemset

Question 3. *Soit le motif séquentiel* $\boldsymbol{p} = \langle d\ \neg(af)\ b\rangle$. *Indiquer les séquences de la base ci-dessous dans lesquelles, selon vous,* $\boldsymbol{p}$ *apparaît.*

id	*Séquence*	
$\boldsymbol{i_0}$	$\langle e\ e\ d\ a\ b\ e\rangle$	☐
$\boldsymbol{i_1}$	$\langle d\ (af) b\ c\rangle$	☐
$\boldsymbol{i_2}$	$\langle e\ d\ (fc)\ b\rangle$	☐
$\boldsymbol{i_3}$	$\langle e\ c\ d\ (ec)\ b\rangle$	☐
$\boldsymbol{i_4}$	$\langle d\ (fa)\ b\ e\rangle$	☐

Cette question est construite de telle sorte que chacune des séquences contienne la partie positive du motif[6] avec seulement un itemset entre les occurrences de d et de b. Ces séquences posent donc la question de la non-inclusion de l'itemset (af) dans a, (af), (fc), (ec) ou (fa). Dans le cas où l'enquêté(e) coche les séquences $\boldsymbol{i_0}$, $\boldsymbol{i_2}$ et $\boldsymbol{i_3}$, on peut en déduire qu'il/elle considère que la présence d'au moins un élément de l'itemset (af) "active" la négation. On parle alors d'une *non-inclusion partielle*. En revanche, si seule la séquence $\boldsymbol{i_3}$ est cochée, alors on peut en déduire qu'il/elle considère qu'il faut la présence de tous les items de l'itemset pour "activer" la négation. On nomme cela une *non-inclusion totale*. En complément, la séquence $\boldsymbol{i_4}$ visait à voir si l'ordre dans l'itemset pouvait importer aux enquêté(e)s et si cela était cohérent avec leur réponse à la séquence $\boldsymbol{p_4}$ de la question 1.

Plus formellement, cette question discrimine entre deux choix d'inclusion entre deux itemsets $P \in 2^{\mathcal{I}} \setminus \{\emptyset\}$ et $I \in 2^{\mathcal{I}}$:

- non-inclusion partielle : $P \not\subseteq_G I \Leftrightarrow \exists e \in P, e \notin I$
- non-inclusion totale : $P \not\subseteq_D I \Leftrightarrow \forall e \in P, e \notin I$

La non-inclusion partielle signifie que $P \setminus I$ est non-vide tant que la non-inclusion totale signifie que P et I sont disjoints (noté D). Dans la suite, le symbole $\not\subseteq_*$ dénote une relation de non-inclusion entre itemsets, indifféremment $\not\subseteq_G$ ou $\not\subseteq_D$.

5.2 Occurrence d'un motif avec négation

Question 4 (Occurrence d'un motif avec négation)**.** *Soit le motif séquentiel* $\boldsymbol{p} = \langle f\ \neg(ea)\ d\rangle$. *Indiquer les séquences de la base ci-dessous dans lesquelles, selon vous,* $\boldsymbol{p}$ *apparaît.*

id	*Séquence*	
$\boldsymbol{e_0}$	$\langle b\ b\ f\ c\ e\ d\ b\rangle$	☐
$\boldsymbol{e_1}$	$\langle b\ f\ e\ a\ c\ b\ d\rangle$	☐
$\boldsymbol{e_2}$	$\langle f\ c\ (ea)\ b\ c\ d\ c\rangle$	☐
$\boldsymbol{e_3}$	$\langle b\ c\ f\ b\ c\ c\ d\rangle$	☐

Dans cette question, la forme du motif séquentiel $\boldsymbol{p} = \langle f\ \neg(ea)\ d\rangle$ est la même à une permutation près des lettres à la précédente et chaque séquence de la base de séquences contient la partie positive de $\boldsymbol{p}$. Mais cette fois, il y a plusieurs itemsets entre les occurrences de f et de d. L'objet de cette question est donc de voir comment cet ensemble d'itemsets positionnés dans la portée de la négation est considéré par un(e) enquêté(e). On estime tout d'abord que, quel que soit l'enquêté(e), $\boldsymbol{p}$ apparaît dans $\boldsymbol{e_3}$ (il n'y a clairement ni e ni a ici) mais $\boldsymbol{p}$ n'apparaît pas dans $\boldsymbol{e_2}$ (on retrouve l'itemset (ea) dans la portée de la négation). La séquence la plus révélatrice est en fait $\boldsymbol{e_1}$. La spécificité de cette séquence est de contenir les deux éléments de l'itemset nié (e et a), mais dans deux itemsets différents. L'enquêté(e) qui ne la coche pas (*i.e.* qu'il/elle considère que $\boldsymbol{p}$ n'apparaît pas dans $\boldsymbol{e_1}$) utilise la notion d'*occurrence souple* : il faudrait que e et a apparaissent ensemble pour "activer" la négation (comme le cas de $\boldsymbol{e_2}$). L'enquêté(e) qui la coche estime que c'est globalement sur toute la période que s'applique la contrainte de négation, on parle d'*occurrence stricte*.

[6] $\langle d\ b\rangle$ est la partie positive de $\langle d\ \neg(af)\ b\rangle$.

Quant à la séquence $\boldsymbol{e_0}$, elle révèle la notion de non-inclusion vu précédemment : en cas de non-inclusion partielle, $\boldsymbol{p}$ apparaît dans $\boldsymbol{e_0}$, mais pas si on considère une non-inclusion totale.

Deux sémantiques ont été distinguées : les occurrences strictes et les occurrences souples. Elles peuvent être formellement définies comme suit : Soit une séquence $\boldsymbol{s} = \langle s_1 \dots s_n \rangle$ et un motif avec négation $\boldsymbol{p} = \langle p_1 \neg q_1 \dots \neg q_{m-1} p_m \rangle$. On dit que $\boldsymbol{e} = (e_i)_{i \in [m]} \in [n]^m$ est une occurrence souple de $\boldsymbol{p}$ dans la séquence $\boldsymbol{s}$ ssi :

- $p_i \subseteq s_{e_i}$ pour tout $i \in [m]$
- $q_i \not\subseteq_* s_j,\ \forall j \in [e_i+1, e_{i+1}-1]$ pour tout $i \in [m-1]$

On dit que $\boldsymbol{e} = (e_i)_{i \in [m]} \in [n]^m$ est une occurrence stricte de $\boldsymbol{p}$ dans la séquence $\boldsymbol{s}$ ssi :

- $p_i \subseteq s_{e_i}$ pour tout $i \in [m]$
- $q_i \not\subseteq_* \bigcup_{j \in [e_i+1, e_{i+1}-1]} s_j$ pour tout $i \in [m-1]$

Intuitivement, la contrainte souple considère la non-inclusion de q_i pour chacun des itemsets situés dans l'intervalle de position $[e_i + 1, e_{i+1} - 1]$ tandis que la contrainte stricte considère la non-inclusion sur l'union de l'ensemble des itemsets à ces mêmes positions. L'intervalle correspond aux itemsets de la séquence strictement entre les occurrences des itemsets entourant q_i.

5.3 Occurrences multiples dans une séquence

Question 5 (Occurrences multiples d'un motif avec négation)**.** *Soit le motif séquentiel $\boldsymbol{p} = \langle b \neg e\ f \rangle$. Indiquer les séquences de la base ci-dessous dans lesquelles, selon vous, $\boldsymbol{p}$ apparaît.*

id	*Séquence*	
$\boldsymbol{o_0}$	$\langle b\ a\ f\ d\ b\ d\ f \rangle$	☐
$\boldsymbol{o_1}$	$\langle b\ a\ f\ d\ e\ b\ d\ f \rangle$	☐
$\boldsymbol{o_2}$	$\langle d\ b\ e\ c\ a\ d\ f\ b\ d\ e\ f \rangle$	☐
$\boldsymbol{o_3}$	$\langle b\ a\ f\ b\ a\ e\ f \rangle$	☐

Dans cette question, les séquences ci-dessous contiennent chacunes plusieurs occurrences de la partie positive du motif. Pour rendre plus visible cette situation, il y a même plusieurs occurrences non imbriquées de $\langle b\ f \rangle$. Dans la mesure où la contrainte de négation porte uniquement sur un seul item (e), les choix relatifs aux dimensions précédentes – non-inclusion d'un itemset et type d'occurrence – n'ont a priori pas d'impact. Ceci permet donc de focaliser la question sur la perception de ces occurrences multiples. Deux comportements sont alors attendus :

- La première interprétation consiste à considérer que dès qu'il existe une occurrence de la partie positive, $\langle b\ f \rangle$, qui satisfait la contrainte de négation, alors la séquence est reconnue. On parle alors d'*occurrence faible*. Cette interprétation est révélée par la sélection des séquences $\boldsymbol{o_0}$, $\boldsymbol{o_1}$ et $\boldsymbol{o_3}$.
- Le second comportement consiste à considérer que dès qu'une occurrence de la partie positive ne satisfait pas la contrainte de négation, alors la séquence n'est pas reconnue. On parle alors de *non-occurrence forte*. Pour la question 5, cela correspond aux enquêté(e)s qui ont coché uniquement la séquence $\boldsymbol{o_0}$, toutes les autres ayant au moins une occurrence de $\langle b\ f \rangle$ avec un e interstitiel. On peut néanmoins constater que la séquence $\boldsymbol{o_1}$ est piégeuse pour ceux qui ont cette intuition, puisqu'il y a deux occurrences minimales de $\langle b\ f \rangle$ (au sens de Mannila et *al.* [9]) qui satisfont la contrainte de négation, mais il y a aussi une occurrence impliquant le premier b et le dernier f. Cette dernière occurrence de la partie positive ne satisfait pas la contrainte de négation. Pour des novices dans l'utilisation des séquences, cette subtilité peut être difficile à détecter. Il semble donc plus judicieux de ne juger de l'interprétation que sur l'absence de $\boldsymbol{o_3}$.

Cette question nous amène de nouveau à deux alternatives. Soit une séquence $\boldsymbol{s}$ et un motif $\boldsymbol{p}$. Pour $\not\subseteq_* \in \{\not\subseteq_D, \not\subseteq_G\}$ et $\bullet \in \{\circ, \bullet\}$,

- $\boldsymbol{p} \preceq^*_\bullet \boldsymbol{s}$ signifie que le motif $\boldsymbol{p}$ est inclus dans la séquence $\boldsymbol{s}$ ssi il existe au moins une occurrence (souple ou stricte) de $\boldsymbol{p}$ dans $\boldsymbol{s}$ avec la non-inclusion $\not\subseteq_*$.
- $\boldsymbol{p} \sqsubseteq^*_\bullet \boldsymbol{s}$ signifie que le motif $\boldsymbol{p}$ est inclus dans la séquence $\boldsymbol{s}$ ssi il existe au moins une occurrence de $\boldsymbol{e}$ dans $\boldsymbol{p}^+$ et que pour chaque occurrence $\boldsymbol{e}$ de $\boldsymbol{p}^+$ dans $\boldsymbol{s}$, $\boldsymbol{e}$ est également une occurrence (souple ou stricte) de $\boldsymbol{p}$ dans $\boldsymbol{s}$ avec la non-inclusion $\not\subseteq_*$.

Les trois dimensions interprétatives de la négation se combinent donc en huit sémantiques possibles définies par leurs relations d'inclusion : $\preceq^D_\circ, \preceq^D_\bullet, \preceq^G_\circ, \preceq^G_\bullet, \sqsubseteq^D_\circ, \sqsubseteq^D_\bullet, \sqsubseteq^G_\circ, \sqsubseteq^G_\bullet$ étudiées dans [2]. Comme illustré, les trois questions ci-dessus ont été construites pour explorer indépendamment chacune des trois dimensions de la sémantique de la négation dans un motif séquentiel. En particulier, nous avons illustré comment la construction des questions permet d'associer, en fonction de la réponse donnée, un(e) enquêté(e) à une sémantique.

6 Analyse et résultats de l'enquête

À l'issue de la période d'enquête, nous avons collecté 124 questionnaires complets. L'expertise auto-estimée dans le domaine de l'extraction de motifs se répartit en 40 novices, 54 ayant des connaissances en science des données et 27 se déclarants familiers avec l'extraction de motifs. 79 se déclarent comme informaticiens, 82 comme chercheurs et 23 comme logiciens. Le nombre de tentatives pour la compréhension de la notion d'occurrence d'un motif est en moyenne de 1.27 ± 0.49 (entre 1 et 5 tentatives). 102 ont correctement répondu dès la première tentative. On peut noter que 6 enquêté(e)s ayant des connaissances en analyse de données (sur 24) ont eu besoin de plus d'une tentative pour avoir la réponse correcte.

Le résultat de l'enquête comporte les réponses booléennes (séquence cochée ou non cochée) pour chaque séquence des questions. Dans l'objectif d'identifier les sémantiques les plus naturelles chez les enquêté(e)s, on peut voir ce problème comme un problème d'extraction d'itemsets fréquents ou de co-clustering. On cherche à identifier des groupes d'individus qui ont coché les mêmes réponses.
Pour l'analyse des réponses, nous procédons en deux temps :

1. on commence par analyser les résultats question par question, *i.e.* indépendamment pour chacune des dimensions de la sémantique des motifs ($\not\subseteq_D$ ou $\not\subseteq_G$, $\circ$ ou $\bullet$, $\preceq$ or $\sqsubseteq$).

2. on complète l'analyse par une analyse globale des sémantiques.

Dans la section précédente, nous avons identifié pour chaque question les grandes classes de réponse attendue. On donne donc par la suite les statistiques d'apparition pour chacune, mais comme les réponses ne correspondent pas forcément exactement à ce qui est attendu (soit par inattention de l'enquêté(e), soit par une interprétation différente), nous proposons d'utiliser l'analyse de concepts formels (*Formal Concept Analysis* ou FCA) [6] pour donner une vision globale des résultats. La FCA est une technique d'analyse de données qui identifie des concepts d'un jeu de données. Chaque concept est décrit, d'une part, par son intention qui est ici un ensemble de réponses cochées et, d'autre part, son extension qui liste tous les individus qui ont choisi ces réponses. Les concepts extraits sont *fermés*, c'est-à-dire que leur extension est maximale pour leur intention et réciproquement. Un des intérêts de la FCA est de représenter de manière synthétique les données dans le treillis de concepts. Au travers de ce treillis, il est possible d'analyser précisément des groupes d'individus ayant fait les mêmes réponses. On peut noter que la FCA a déjà été utilisée pour l'analyse de questionnaires [1]. L'outil utilisé pour construite les treillis est GALACTIC [3].

6.1 Analyse de chaque dimension de la sémantique

Dans cette partie, on analyse les réponses à quatre questions : on s'intéresse tout d'abord aux réponses à la question sur la portée des négations, ensuite, on analyse les trois dimensions de la sémantique des motifs avec négation : la non-inclusion des itemsets, les occurrences et les relations d'inclusion. Les Tableaux 2 à 5 donnent de manière synthétique les nombres de chacune des interprétations. Les Figures 2 à 4 illustrent les treillis de concepts obtenus pour chacune de ces questions pour donner une image plus globale des réponses.
Concernant la compréhension de la portée des négations, 101 personnes ont coché des réponses correspondant à l'attendu pour cette question de vérification (cf. Table 2). Il est intéressant de constater que 9 personnes qui avaient coché $\boldsymbol{s_1}$ et $\boldsymbol{s_3}$ n'ont pas coché $\boldsymbol{s_4}$ laissant penser que, pour elles, la négation d'un itemset signifie qu'il s'agit d'un

Table 2: Résultat sur la question de la portée de la négation.

Portée	**Nombre**	**Pourcentage**
Conforme	101	81.4%
Conforme sauf $\boldsymbol{s_4}$	9	7.3%
Alternatif	14	11.3%

Table 3: Réponses à la question des non-inclusions (en nombre et en pourcentage).

Interprétation	**Nombre**	**Pourcentage**
Non-inclusion partielle	100	90.9%
Non-inclusion totale	3	2.7%
Autre	7	6.4%

évènement qui n'est pas l'évènement nié. Autrement dit, il faut au moins un évènement qui ne soit pas l'évènement nié pour activer la négation.[7] Pour les autres différences marginales (14 personnes), nous considérons qu'il s'agit d'oublis ou d'erreurs. Les réponses de ces personnes ont été écartées de la suite de l'analyse des résultats, leur compréhension possiblement différente de la portée de la négation ne permet d'exploiter leurs réponses. La suite des analyses porte donc sur 110 questionnaires complétés.
Concernant les non-inclusions d'itemsets (Table 3 et Figure 2), on constate que les enquêté(e)s ont très majoritairement (100) sélectionné le triplet de réponse $\boldsymbol{i_0}$, $\boldsymbol{i_2}$ et $\boldsymbol{i_3}$ correspondant à l'interprétation de non-inclusion partielle (concept `$8` dans la Figure 2). Seulement 3 personnes ont considéré l'interprétation de la non-inclusion totale. De manière plutôt inattendue, 22 enquêté(e)s ont considéré que la séquence $\boldsymbol{i_4}$ contenait le motif et donc que (fa) n'était pas incompatible avec (af). Ces enquêté(e)s se répartissent dans les différents niveaux d'expertises (8, 11 et 3 respectivement pour les niveaux 0, 1 et 2). Il ne s'agit donc pas plus particulièrement des personnes qui ne sont pas biaisées par l'habitude des notations de la fouille de motifs.

Table 4: Réponses à la question des occurrences.

Interprétation	**Nombre**	**Pourcentage**
Occurrence stricte	97	88.2%
Occurrence souple	7	6.3%
Autre	6	5.5%

Concernant l'analyse des occurrences (Table 4 et Figure 3), seule la séquence $\boldsymbol{e_1}$ permet de discriminer l'intuition des enquêté(e)s. Pour la Table 4, on s'assure aussi que les réponses sont correctes pour $\boldsymbol{e_2}$ et $\boldsymbol{e_3}$, sinon on place la réponse en "autre". De nouveau, on obtient un résultat très marqué pour l'interprétation dite souple : 97 personnes y adhèrent (concept `$7` dans la Figure 3). Le concept `$3` correspond aux individus qui n'ont pas coché $\boldsymbol{e_1}$. Il s'agit donc des interprétations d'occurrence stricte.

[7] NB : dans les questions suivantes, toutes les séquences ont au moins un évènement "neutre" là où un itemset avec négation est attendu. On peut donc conserver des personnes sans biaiser les réponses suivantes.

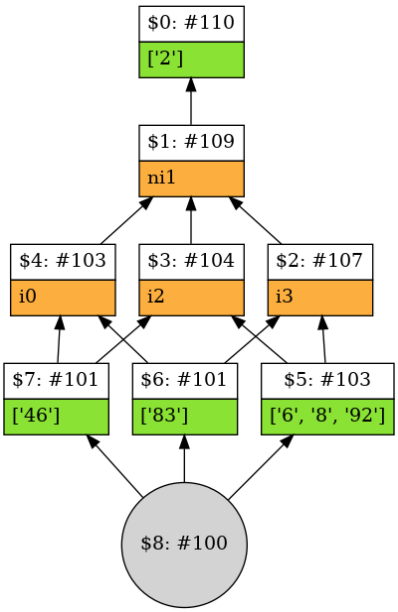

Figure 2: Concepts extraits à partir des réponses à la Question 3 : non-inclusion d'un itemset. Chaque concept est illustré par une boîte contenant différents éléments : les générateurs sur fond orange (réponses possibles aux questions), les prototypes sur fond vert. Le symbole `$` désigne un numéro de concept (un identifiant). La taille de l'extension est précisée avec un `#`. La liste entre crochets pour un prototype désigne une liste d'exemples (ici des numéros de questionnaires) "complémentaire" par rapport au concept inférieur. Par exemple, le concept `$7` couvre 101 exemples: les 100 du concept `$8` plus l'exemple `46`. Chaque concept indique l'intention comme un ensemble de séquences cochées (se reporter aux tables présentées dans les exemples). Dans les réponses aux questions, `i0` désigne que d'enquêté(e) a coché la séquence $\boldsymbol{i_0}$, et `ni1` (préfixe avec `n`) désigne que d'enquêté(e) **n'a pas** coché la séquence $\boldsymbol{i_1}$. Le choix entre les deux représentations d'une réponse a été fait en considérant la lisibilité du treillis obtenu.

Finalement, concernant l'analyse des relations d'inclusion (Table 5 et Figure 4), le résultat est ici plus partagé. 75 personnes ont exclusivement identifié les trois séquences correspondant à la notion de relation faible (relation $\preceq$). Ils sont représentés dans le concept `$3` de la Figure 4. En revanche, 31 personnes ont exclusivement sélectionné la séquence $\boldsymbol{o_0}$ (concept `$1`). Ces derniers ont ainsi préféré l'interprétation de la relation forte (relation $\sqsubseteq$). Ces 31 personnes comprennent 14 qui n'ont pas coché la séquence $\boldsymbol{o_1}$ et 17 qui l'ont cochée. Ces derniers adhèrent plus à la notion d'occurrence minimale de Mannila et *al.* [10].

6.2 Fréquence des sémantiques

Les questions 3, 4 et 5 attribuent chaque enquêté(e) à une interprétation d'une des trois dimensions qui constituent la sémantique d'un motif avec négation. On cherche maintenant à voir s'il existe des sémantiques (comme combinaison des choix d'interprétation pour les trois dimensions) qui sont dominantes parmi les 8 possibles.

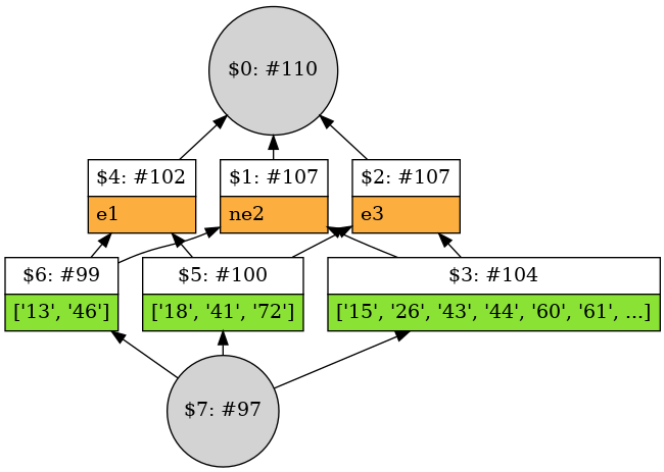

Figure 3: Concepts extraits à partir des réponses à la Question 4 relative aux occurrences (cf Figure 2 pour légende).

Table 5: Réponses à la question des relations d'inclusion.

Interprétation	Nombre	Pourcentage
Relation faible	75	69.2%
Relation forte	31	28.2%
Autre	4	3.6%

La Figure 5 synthétise les réponses à notre enquête. Elle illustre le treillis de concepts représentant les sémantiques des motifs avec négation. Les cinq prototypes du niveau inférieur décrivent les 5 sémantiques (et leur représentation dans les données) qui ont été effectivement utilisées par les enquêté(e)s. Ils sont définis par un choix d'interprétation pour chacune des trois dimensions. De gauche à droite, on identifie les sémantiques $\preceq^D_\circ$, $\preceq^G_\bullet$, $\sqsubseteq^G_\bullet$, $\sqsubseteq^G_\circ$ et $\sqsubseteq^D_\circ$. Sachant que parmi les huit sémantiques, il y a en fait deux paires équivalentes [2] ($\sqsubseteq^D_\circ \sim \sqsubseteq^D_\bullet$ et $\preceq^D_\circ \sim \preceq^D_\bullet$), la seule sémantique qui n'est pas représentée dans cette enquête est $\preceq^G_\circ$.

Sur les 110 enquêté(e)s, le questionnaire a permis d'attribuer une sémantique intuitive à 96 personnes. Les 14 autres personnes ont au moins une question pour laquelle il n'a pas été clairement identifiée une des interprétations attendues. Ces individus se trouvent dans les concepts intermédiaires (prototypes `5`, `6` et `10`, générateur `15` et concepts `3`, `9` et `13`).

Une première constatation est qu'il est possible d'attribuer une sémantique à une grande partie des enquêté(e)s. C'est-à-dire que ce sont probablement les mêmes personnes qui ont faits des réponses "alternatives" aux différentes questions. Ce résultat nous conforte sur l'exploitabilité des résultats collectés.

Ensuite, cette figure met en évidence le résultat principal de cette étude : il existe principalement deux sémantiques qui sont intuitivement utilisées : $\sqsubseteq^G_\circ$ à 23.9% et $\preceq^G_\circ$ à 69.8%. Les autres sémantiques sont marginalement représentées.

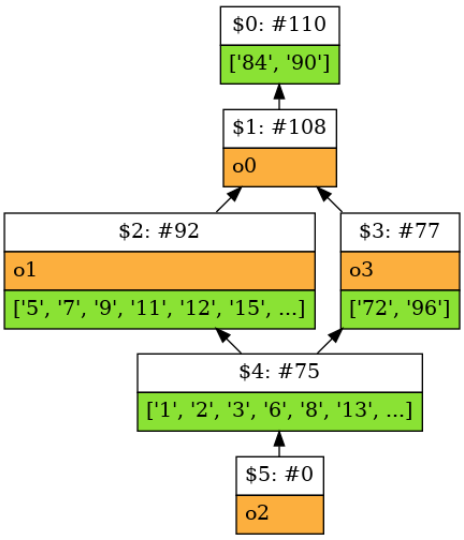

Figure 4: Concepts extraits à partir des réponses à la Question 5 relative aux relations d'inclusion (cf. Figure 2 pour légende).

Nous nous sommes ensuite intéressés à la comparaison des populations définies par le choix des sémantiques en nous intéressant à leurs réponses aux questions sur leur profil. Nous avons pour cela procédé à un test statistique pour comparer les distributions des niveaux d'expertises (test de Student). Les résultats ne montrent aucune différence significative entre les groupes. La conclusion que nous tirons est que l'intuition d'une sémantique n'est globalement pas liée à une expertise particulière en informatique ou en science des données.

7 Sémantiques préférées et algorithmes

On peut conclure de ces analyses qu'il n'y a pas une seule sémantique partagée pour les enquêté(e)s, mais plutôt que deux dominent : $\sqsubseteq_{\circ}^{G}$ et $\preceq_{\circ}^{G}$. Il est intéressant de comparer ce résultat avec les choix des deux algorithmes majeurs du domaine, eNSP et NEGPSPAN dont les sémantiques sont respectivement $\sqsubseteq_{\circ}^{D}$ et $\preceq_{\circ}^{D}$/$\preceq_{\bullet}^{D}$.

Tout d'abord, aucun des algorithmes ne répond à l'intuition des enquêté(e)s puisque les deux s'appuient sur une non-inclusion totale des itemsets tandis que c'est la non-inclusion partielle qui semble la plus intuitive. Une explication du choix algorithmique vient du fait que la non-inclusion partielle est antimonotone tandis que la relation totale est monotone. Cette dernière est moins facile à exploiter algorithmiquement. Les sémantiques les plus intuitives ne sont donc pas celles qui sont les plus appropriées algorithmiquement.

En pratique, on peut craindre des erreurs d'interprétation des motifs extraits par les algorithmes. Sans explicitation de la sémantique de ces derniers, les résultats de cette étude montrent que les motifs seront interprétés avec une sémantique différente de celle qui a servi à les extraire. Ceci constitue donc un problème important sur l'utilisation de ces algorithmes.

Une première recommandation serait alors de n'avoir que des singletons dans les négations d'un motif. Auquel cas, les non-inclusions partielles et totales sont équivalentes.

Une seconde solution serait de développer un algorithme alternatif adapté à une interprétation partielle de la non-inclusion. Ces adaptations sont algorithmiquement faisables. Il faudrait alors comparer leurs performances de calcul pour s'assurer que de tels algorithmes restent efficaces.

Néanmoins, les résultats montrent que le choix effectué par NEGPSPAN concernant la gestion des occurrences multiples répond à l'intuition d'un plus grand nombre. La seconde recommandation serait donc d'étendre préférentiellement l'algorithme NEGPSPAN.

Finalement, la troisième recommandation serait de promouvoir l'utilisation de syntaxes différenciées pour chaque sémantique. Cette recommandation avait été également suggérée dans [2].

8 Discussion

Cette enquête est probablement perfectible sur sa méthodologie. En particulier, il y a eu peu de questions pour décrire précisément le profil des enquêté(e)s. Ceci ne permet pas de savoir si la population enquêtée correspond bien à celle des utilisateurs potentiels d'algorithmes de fouille de motifs. De plus, la diffusion du questionnaire a été majoritairement effectuée via des canaux académiques. Ceci peut introduire un biais dans les réponses.

Une seconde limite sur la forme du questionnaire est la non-redondance des questions. En effet, chaque dimension d'interprétation de la sémantique des motifs avec négation ne fait l'objet que d'une seule question. Cela peut être jugé sensible à des erreurs. Nous avons opté pour un questionnaire plus court ne répétant par les questions. De plus, nous avons conçu le questionnaire pour séparer au mieux les différentes dimensions et ainsi éviter toute ambiguïté dans l'analyse des résultats.

La troisième limite est que le nombre de réponse au questionnaire peut sembler faible. La collecte de 124 questionnaires a nécessité plusieurs mois. La récupération d'un nombre significativement supérieur aurait nécessité d'autres stratégies de diffusion. De plus, ce nombre nous est apparu suffisant, au regard des questions et des résultats, pour faire une analyse solide statistiquement. En effet, les différences marquées dans les résultats permettent de fournir des résultats significatifs.

La qualité des réponses collectées est attestée par deux questions : une question préliminaire éliminatoire ainsi qu'une seconde question portant sur la portée de la négation qui nous sert à filtrer les enquêté(e)s qui viendraient biaiser les résultats. Le très faible nombre de ces personnes laissent penser que le jeu de réponses est de bonne qualité, *i.e.* que les personnes enquêtées ont consciencieusement répondu aux questions. Ceci est conforté par le fait qu'on ait pu attribuer une sémantique à 88% des enquêté(e)s

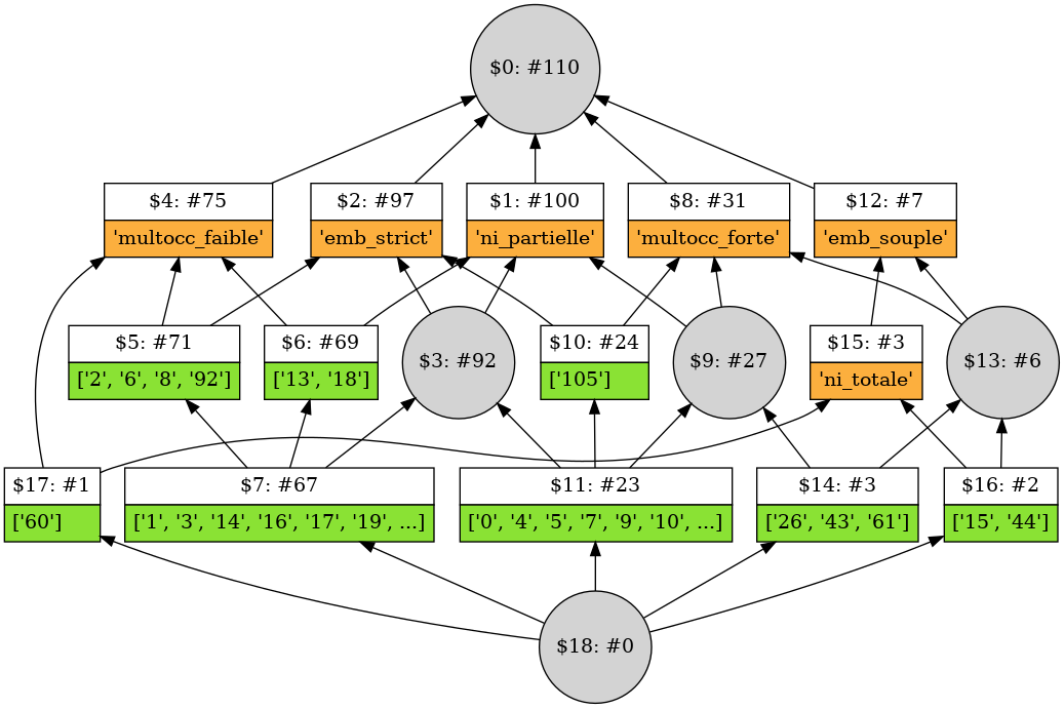

Figure 5: Concepts extraits à partir des réponses attributions faites pour chaque dimension.

montrant que les interprétations alternatives ou erreurs sont concentrées sur un petit nombre de personnes.

Un autre biais possiblement important de ce questionnaire est la présentation des notions élémentaires de motifs séquentiels qui pourraient avoir induit certaines réponses plutôt que d'autres. On peut en effet s'interroger sur le fait que les réponses pour les questions 3 et 4 aient des réponses aussi peu diverses. On s'attendait à avoir des perceptions plus hétérogènes de la notion de non-inclusion d'itemsets, mais cette diversité ne se retrouve pas dans le panel de personnes interrogées. Dans la mesure où la question 5 montre de la diversité dans les réponses, nous estimons que si l'hétérogénéité était réellement marquée dans les questions précédentes, elle serait apparue dans les réponses au questionnaire. Parmi les biais de présentation, la forme de présentation du questionnaire avec des symboles (et non des lettres) nous a été rapportée comme intéressante par certain(e)s enquêté(e)s. En effet, l'utilisation de lettres présuppose un ordre dans les items qui n'existe pas. En pratique, on a observé que seul 22.6% des enquêté(e)s étaient sensibles à l'ordre dans les itemsets. L'utilisation de symboles géométriques retranscrit mieux cette idée d'*itemset*. Malheureusement, nous n'avons pas collecté l'information sur le mode graphique du questionnaire effectivement utilisé. Nous ne pouvons donc pas vérifier cette hypothèse.

Finalement, le questionnaire est intimement lié au cadre d'analyse proposé par Besnard et Guyet [2] qui fait quelques hypothèses sur la forme des motifs avec négation, et leur sémantique. En particulier, nous avons vu ci-dessus que la question de l'insensibilité à l'ordre dans un itemset est une hypothèse forte que nous faisons. La seconde hypothèse est sur la portée de la négation. On s'aperçoit que 18.5% des enquêté(e)s n'ont pas répondu comme attendu à cette question. Comme nous avons écarté ces personnes de l'analyse, cela n'impacte pas les conclusions, mais cela soulève des questions sur l'"intuition" qu'ont eu ces personnes. Des interviews plus précises seraient ici nécessaires. Une troisième hypothèse est sur la syntaxe des motifs avec négation. Une étude plus complète pourrait s'intéresser à des syntaxes plus étendues : en permettant, par exemple, plusieurs négations consécutives, ou des négations en tête ou en queue d'un motif. Ces dernières possibilités existent chez certains algorithmes d'extraction de motifs de l'état de l'art [8].

9 Conclusion

Dans cet article, nous nous sommes intéressés à la sémantique des motifs séquentiels avec négation du point de vue des utilisateur(trice)s potentiels d'algorithmes d'extraction de tels motifs. L'intérêt de ce travail est de savoir si les motifs qui sont extraits par les algorithmes de l'état de l'art sont bien interprétés par les utilisateur(trice)s. En effet, les travaux de l'état de l'art avaient mis en évidence une ambiguïté dans ces notations [2]. Pour répondre à cette question, nous avons mené une enquête auprès d'utilisateur(trice)s potentiels ayant des profils variés. Cette enquête visait à comprendre les sémantiques auxquelles les utilisateurs adhéraient plus favorablement parmi celles qui avaient été identifiées.

L'analyse des réponses à l'enquête montre que deux sémantiques, dénotées $\sqsubseteq_{\circ}^{G}$ et $\preceq_{\circ}^{G}$, dominent dans le panel de 124 personnes interrogées. Il est tout d'abord intéressant

de constater qu'il n'existe pas une sémantique intuitive partagée uniformément. Les résultats sont également particulièrement intéressants par le fait que la préférence pour $\not\subseteq_G$ ne correspond pas à ce qui est utilisé dans les algorithmes majeurs de l'extraction de motifs avec négation (eNSP et NEGPSPAN). Cette relation intervenant lorsque la négation porte sur des ensembles d'item (*i.e.* $\neg(ab)$), une information particulière devrait être donnée aux utilisateurs sur les motifs comportant ce type de contrainte. Ensuite, $\preceq$ est majoritaire (à $\approx 69\%$) dans le panel et correspond au choix de l'algorithme NEGPSPAN. C'est également à la sémantique qui dispose des propriétés d'antimonotonie si les négations ne portent que sur des singletons.
À la suite de cette enquête, nous formulons les recommandations suivantes pour les méthodes d'extraction de motifs séquentiels avec négation :

- limiter l'usage de négation d'itemsets et privilégier l'utilisation de négation d'items,
- sinon, explorer l'extension l'algorithme NEGPSPAN dont la sémantique de la relation d'inclusion correspond à l'intuition majoritaire,
- proposer des syntaxes différenciées pour les différentes sémantiques.

References

[1] Radim Belohlavek, Erik Sigmund, and Jiří Zacpal. Evaluation of IPAQ questionnaires supported by formal concept analysis. *Information Sciences*, 181(10):1774–1786, 2011.

[2] Philippe Besnard and Thomas Guyet. Semantics of negative sequential patterns. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 1009–1015. IOS Press, 2020.

[3] Salah Eddine Boukhetta, Christophe Demko, Karell Bertet, Jérémy Richard, and Cécile Cayèré. Temporal sequence mining using FCA and GALACTIC. In *Graph-Based Representation and Reasoning*, pages 185–199. Springer International Publishing, 2021.

[4] Longbing Cao, Xiangjun Dong, and Zhigang Zheng. e-NSP: Efficient negative sequential pattern mining. *Artificial Intelligence*, 235:156–182, 2016.

[5] Xiangjun Dong, Yongshun Gong, and Longbing Cao. e-RNSP: An efficient method for mining repetition negative sequential patterns. *IEEE transactions on cybernetics*, 50(5):2084–2096, 2018.

[6] Bernhard Ganter and Rudolf Wille. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media, 2012.

[7] Thomas Guyet and René Quiniou. NegPSpan: efficient extraction of negative sequential patterns with embedding constraints. *Data Mining and Knowledge Discovery*, 34(2):563–609, 2020.

[8] Sue-Chen Hsueh, Ming-Yen Lin, and Chien-Liang Chen. Mining negative sequential patterns for e-commerce recommendations. In *Proceedings of the Asia-Pacific Services Computing Conference*, pages 1213–1218. IEEE, 2008.

[9] Heikki Mannila and Hannu Toivonen. Discovering generalized episodes using minimal occurrences. In *Proceedings of the Conference on Knowledge Discovery and Delivery (KDD)*, volume 96, pages 146–151, 1996.

[10] Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data mining and knowledge discovery*, 1(3):259–289, 1997.

[11] A. Savasere, E. Omiecinski, and S. Navathe. Mining for strong negative associations in a large database of customer transactions. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 494–502, 1998.

[12] Wei Wang and Longbing Cao. Negative sequence analysis: A review. *ACM Computing Survey*, 52(2):32:1–32:39, 2019.

[13] Wei Wang and Longbing Cao. VM-NSP: vertical negative sequential pattern mining with loose negative element constraints. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–27, 2021.

[14] Xindong Wu, Chengqi Zhang, and Shichao Zhang. Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems (TOIS)*, 22(3):381–405, 2004.

[15] Tiantian Xu, Xiangjun Dong, Jianliang Xu, and Xue Dong. Mining high utility sequential patterns with negative item values. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 31(10):1750035, 2017.

[16] Tiantian Xu, Xiangjun Dong, Jianliang Xu, and Yongshun Gong. E-msNSP: Efficient negative sequential patterns mining based on multiple minimum supports. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 31(02):1750003, 2017.