24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Semantics of negative sequential patterns

Philippe Besnard [1] and **Guyet Thomas** [2]

Abstract. In the field of pattern mining, a negative sequential pattern is specified by means of a sequence consisting of events to occur and of other events, called negative events, to be absent. For instance, containment of the pattern $\langle a\ \neg b\ c\rangle$ arises with an occurrence of a and a subsequent occurrence of c but no occurrence of b in between.

This article is to shed light on the ambiguity of such a seemingly intuitive notation and we identify eight possible semantics for the containment relation between a pattern and a sequence. These semantics are illustrated and formally studied, in particular we propose dominance and equivalence relations between them. Also we prove that support is anti-monotonic for some of these semantics. Some of the results are discussed with the aim of developing algorithms to extract efficiently frequent negative patterns.

1 Introduction

In many application domains such as predictive maintenance or marketing, decision makers are interested in discovering specific events that trigger or are correlated to undesirable events. Sequential pattern mining [11] is a technique that extracts such hidden rules from logs.

Often, the presence but also the absence of a specific action or event partly explains the occurrence of an undesirable situation [4]. For example in predictive maintenance, if some maintenance operations have not been performed, *e.g.* damaged parts have not been replaced, then a fault is likely to occur in a short delay whereas if these operations were performed in time the fault would not occur. In marketing, if a marketplace customer has not received special offers or coupons for a long time then s/he has a high probability of churning whereas if s/he were provided with such offers s/he should remain loyal to her/his marketplace. Mining specific events to discover a context under which they occur, or do not occur, may provide interesting information. It is called *actionable* information as it serves to determine what action should be performed to avoid the undesirable situation, *i.e.* fault in monitored systems, churn in marketing, . . .

Figure 1. On the left, a synthetic dataset of six sequences. On the right, a set of rules with their support and accuracy in the dataset.

s_1	$\langle a\ c\ e\rangle$
s_2	$\langle a\ b\ c\ e\rangle$
s_3	$\langle a\ b\ c\ e\rangle$
s_4	$\langle a\ c\ d\rangle$
s_5	$\langle a\ c\ d\rangle$
s_6	$\langle a\ c\ d\rangle$

Rule	support	accuracy
$\langle a\ c\rangle \implies e$	3	$^1/_2$
$\langle a\ c\rangle \implies d$	3	$^1/_2$
$\langle a\ \neg b\ c\rangle \implies e$	1	$^1/_4$
$\langle a\ \neg b\ c\rangle \implies d$	3	$^3/_4$

Standard sequential pattern mining algorithms [11] extract sequential patterns that frequently occur in the logs. A sequential pattern is a sequence of events. For example, the sequential pattern $\langle a\ c\ d\rangle$ is read as "a occurs and then c occurs and finally d occurs". In practice, a pattern is frequent if its number of occurrences exceeds a user-defined threshold. If $\langle a\ c\ d\rangle$ occurs in most cases where $\langle a\ c\rangle$ occurs, then the sequential rule $\langle a\ c\rangle \implies d$ (read as "if a occurs and c occurs later, then d occurs afterwards") is useful to predict occurrences of d. The premise of a rule specifies what actually occurred frequently, but does not inform about what did not happen in these examples. Negative sequential patterns are sequential patterns that also specify non-occurring events. Intuitively, the syntax of a simple negative sequential pattern is as follows: $\langle a\ \neg b\ c\rangle$. This pattern is read as "a occurs and then c occurs, but b does not occur in between". A negative sequential pattern can also be the premise of a rule.

We illustrate the interest of negative sequential patterns via the dataset of sequences in Figure 1. The rightmost table gives the support (number of sequences containing both premise and conclusion) and the accuracy (ratio of the support with the support of the premise only) of some rules. The sequential patterns $\langle a\ c\ d\rangle$ and $\langle a\ c\ e\rangle$ occur thrice each. Rules $\langle a\ c\rangle \Rightarrow d$ and $\langle a\ c\rangle \Rightarrow e$ obtained from positive sequential patterns have low accuracy, they are not really interesting. Let $\boldsymbol{s_p} = \langle a\ b\ c\ ?\rangle$ be a new sequence with an event to predict. The two rules above predict e or d with the same likelihood.

Modeling the absence of event b in patterns appears to be meaningful to describe the dataset. Indeed, rule $\langle a\ \neg b\ c\rangle \Rightarrow d$ occurs in half of the sequences and has an accuracy of $^3/_4$ (whereas the accuracy of the rule without $\neg b$ is only $^1/_2$). When it comes to predicting occurrences of d, the absence of b is meaningful. These new rules predict event e with a likelihood of $^3/_4$ for $\boldsymbol{s_p}$. In a medical context, a, b and c may be drug administration while d and e some medical events, respectively, patient declared cured and patient suffering complications. The situation that is illustrated by our synthetic dataset is the case of adverse drugs reaction. Being exposed to drug b while being treated by drugs a and c leads to complications. Mining positive patterns in a medical database would miss such adverse drug reaction.

Mining frequent negative sequential patterns is of utmost interest to discover actionable rules taking into account absent events. In [10], pattern mining is viewed as the computation of a theory $Th(\mathcal{L}, \mathcal{D}, C) = \{\psi \in \mathcal{L} \mid C(\psi, \mathcal{D})\}$. Given a pattern language $\mathcal{L}$, some constraints C and a database $\mathcal{D}$, a pattern mining algorithm enumerates the elements of the language that fulfill the constraints within the data. In the case of frequent pattern mining, C is the minimal support constraint. The success of pattern mining techniques comes from an anti-monotonicity property of some support measures [1]. Intuitively, if a pattern $\boldsymbol{p}$ is not frequent, no pattern "larger" than $\boldsymbol{p}$ is frequent. Pattern mining algorithms prune the search space whenever an unfrequent pattern is found. The "is larger than" relation induces a partial order on the set of patterns, $\mathcal{L}$. For a support measure that is anti-

[1] CNRS / IRIT, France, email: besnard@irit.fr
[2] Institut Agro / IRISA UMR6074, France, email: thomas.guyet@irisa.fr

monotonic on this structure, the frequent pattern mining trick can be used to efficiently prune the search space. Ideally, this structure is a lattice, in which case the above strategy is complete and correct.

As to frequent negative sequential pattern mining, $\mathcal{L}$ is the set of negative sequential patterns, $\mathcal{D}$ is a dataset of sequences and C is the constraint of minimal frequency. Few approaches [3, 5, 7, 8, 9, 14, 15] proposed algorithms to extract such patterns and none of them proposed an algorithm based on an anti-monotonic support measure. The questions we address in this article are:

1. what is a proper support measure for negative sequential patterns?
2. is there a support measure enjoying anti-monotonicity?

The support measure is strongly related to the *containment relation* that determines whether a pattern occurs in a sequence or not. In the case of negative sequential patterns, the apparently intuitive notion of absent event appears to be intricate and the negation syntax (the $\neg$ symbol) used in the literature is hiding different semantics. In logic, it is accepted knowledge that there is more than one kind of negation [13]. For instance, in classical reasoning $\neg p$ means that p is false while in stable reasoning $\neg p$ means that p cannot be proved [2].

The objective of this article is not to propose a new pattern mining algorithm for negative sequential patterns but to establish formal results on containment relations that can serve as a basis to design such algorithms. The main contributions of our work are as follows:

- we define eight possible semantics for the containment relation of negative sequential patterns,
- we establish dominance and equivalence relations between containment relations,
- we provide three partial orders for which some containment relations induce anti-monotonic support measures.

2 Negative sequential patterns

Throughout this article, $[n] = \{1, \ldots, n\}$ denotes the set of the first n positive integers. Let $\mathcal{I}$ be the set of items (alphabet). An *itemset* $A = \{a_1\ a_2\ \cdots\ a_m\} \subseteq \mathcal{I}$ is a finite set of items. The length of A, denoted $|A|$, is m. A *sequence* $\boldsymbol{s}$ is of the form $\boldsymbol{s} = \langle s_1\ s_2\ \cdots\ s_n\rangle$ where each s_i is an itemset.

Definition 1 (Negative sequential patterns (NSP))**.** *A negative sequential pattern* $\boldsymbol{p} = \langle p_1\ \neg q_1\ p_2\ \neg q_2\ \cdots\ p_{n-1}\ \neg q_{n-1}\ p_n\rangle$ *is a finite sequence where* $p_i \in 2^{\mathcal{I}} \setminus \{\emptyset\}$ *for all* $i \in [n]$ *and* $q_i \in 2^{\mathcal{I}}$ *for all* $i \in [n-1]$.

The length *of* $\boldsymbol{p}$*, denoted* $|\boldsymbol{p}|$ *is the number of its non empty itemsets (negative or positive).*

$\boldsymbol{p}^+ = \langle p_1\ \ldots\ p_n\rangle$ *is called the positive part of the NSP.*

We denote by $\mathcal{N}$ the set of negative sequential patterns.

It can be noticed that Definition 1 introduces syntactic limitations on negative sequential patterns that are commonly encountered in the state of the art [12]:

- a pattern can neither start or finish by a negative itemset,
- a pattern cannot have two successive negative itemsets.

Example 1 (Negative sequential pattern)**.** *This example illustrates the notations introduced in Definition 1. Consider* $\mathcal{I} = \{a, b, c, d\}$ *and* $\boldsymbol{p} = \langle a\ \neg(bc)\ (ad)\ d\ \neg(ab)\ d\rangle$*. Let* $p_1 = \{a\}$*,* $p_2 = \{ad\}$*,* $p_3 = \{d\}$*,* $p_4 = \{d\}$ *and* $q_1 = \{bc\}$*,* $q_2 = \emptyset$*,* $q_3 = \{ab\}$*. The length of* $\boldsymbol{p}$ *is* $|\boldsymbol{p}| = 6$ *and* $\boldsymbol{p}^+ = \langle a\ (ad)\ d\ d\rangle$*.*

3 Semantics of negative sequential patterns

The semantics of negative sequential patterns relies upon *negative containment*: a sequence $\boldsymbol{s}$ supports pattern $\boldsymbol{p}$ (or $\boldsymbol{p}$ matches the sequence $\boldsymbol{s}$) iff $\boldsymbol{s}$ contains a sub-sequence $\boldsymbol{s}'$ such that every positive itemset of $\boldsymbol{p}$ is included in some itemset of $\boldsymbol{s}'$ in the same order and for any negative itemset $\neg q_i$ of $\boldsymbol{p}$, q_i is *not included* in any itemset occurring in the sub-sequence of $\boldsymbol{s}'$ located between the occurrence of the positive itemset preceding $\neg q_i$ in $\boldsymbol{p}$ and the occurrence of the positive itemset following $\neg q_i$ in $\boldsymbol{p}$.

Definition 2 (Non inclusion)**.** *We introduce two relations comparing two itemsets* $P \in 2^{\mathcal{I}} \setminus \{\emptyset\}$ *and* $I \in 2^{\mathcal{I}}$*:*

- *partial non inclusion:* $P \nsubseteq_G I \Leftrightarrow \exists e \in P,\ e \notin I$
- *total non inclusion:* $P \nsubseteq_D I \Leftrightarrow \forall e \in P, e \notin I$

Partial non-inclusion means that $P \setminus I$ *is non-empty while total non-inclusion means that* P *and* I *are disjoint. By convention,* $\emptyset \nsubseteq_D I$ *and* $\emptyset \nsubseteq_G I$ *for all* $I \subseteq \mathcal{I}$*.*

In the sequel we will denote the general form of itemset non-inclusion by the symbol $\nsubseteq_*$, meaning either $\nsubseteq_G$ or $\nsubseteq_D$.

Intuitively, partial non-inclusion identifies the itemset P with a disjunction of negative constraints, *i.e.* at least one of the items (of P) has to be absent from I, and total non-inclusion consider the itemset P as a conjunction of negative constraints: all items (of P) have to be absent from I.

Choosing one non-inclusion interpretation or the other has consequences on extracted patterns as well as on pattern search. Let us illustrate this with the following dataset of sequences:

$$\mathcal{D} = \left\{ \begin{array}{l} \boldsymbol{s}_1 = \langle (bc)\ f\ a\rangle \\ \boldsymbol{s}_2 = \langle (bc)\ (cf)\ a\rangle \\ \boldsymbol{s}_3 = \langle (bc)\ (df)\ a\rangle \\ \boldsymbol{s}_4 = \langle (bc)\ (ef)\ a\rangle \\ \boldsymbol{s}_5 = \langle (bc)\ (cdef)\ a\rangle \end{array} \right\}.$$

Table 1 compares the support of patterns under the two semantics of itemset non-inclusion. Since the positive part of $\boldsymbol{p}_2$ is in $\boldsymbol{s}_2$, $\boldsymbol{p}_2$ occurs in the sequence iff $(cd) \nsubseteq_* (cf)$. As for total non-inclusion, it is false that $(cd) \nsubseteq_D (cf)$ because c occurs in (cf), and thus $\boldsymbol{p}_2$ does not occur in $\boldsymbol{s}_2$. As for partial non-inclusion, it is true that $(cd) \nsubseteq_G (cf)$, because d does not occur in (cf), and thus $\boldsymbol{p}_2$ occurs in $\boldsymbol{s}_2$.

Lemma 1. [3] *Let* $P, I \subseteq \mathcal{I}$ *be two itemsets:*

$$P \nsubseteq_D I \implies P \nsubseteq_G I \qquad (1)$$

Table 1. Lists of sequences in $\mathcal{D}$ supported by negative patterns $(\boldsymbol{p}_i)_{i=1..4}$ under the total and partial non-inclusion relations. Each pattern has the form $\langle b\ \neg q_i\ a\rangle$ where q_i are itemsets such that $q_i \subset q_{i+1}$.

	partial non-inclusion $\nsubseteq_G$	total non-inclusion $\nsubseteq_D$
$\boldsymbol{p}_1 = \langle b\ \neg c\ a\rangle$	$\{\boldsymbol{s}_1, \boldsymbol{s}_3, \boldsymbol{s}_4\}$	$\{\boldsymbol{s}_1, \boldsymbol{s}_3, \boldsymbol{s}_4\}$
$\boldsymbol{p}_2 = \langle b\ \neg(cd)\ a\rangle$	$\{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4\}$	$\{\boldsymbol{s}_1, \boldsymbol{s}_4\}$
$\boldsymbol{p}_3 = \langle b\ \neg(cde)\ a\rangle$	$\{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4\}$	$\{\boldsymbol{s}_1\}$
$\boldsymbol{p}_4 = \langle b\ \neg(cdeg)\ a\rangle$	$\{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4, \boldsymbol{s}_5\}$	$\{\boldsymbol{s}_1\}$

Now, we formulate the notions of sub-sequence, non-inclusion and absence by means of the concept of embedding.

[3] All proofs can be found at the following link: `https://hal.inria.fr/hal-02481240`.

Definition 3 (Positive pattern embedding)**.** *Let* $\boldsymbol{s} = \langle s_1 \ldots s_n \rangle$ *be a sequence and* $\boldsymbol{p} = \langle p_1 \ldots p_m \rangle$ *be a (positive) sequential pattern. A tuple* $\boldsymbol{e} = (e_i)_{i \in [m]} \in [n]^m$ *is an* embedding *of pattern* $\boldsymbol{p}$ *in sequence* $\boldsymbol{s}$ *iff* $\forall i \in [m],\ p_i \subseteq s_{e_i}$ *and* $e_i < e_{i+1}$ *for all* $i \in [m-1]$.

Definition 4 (Strict and soft embeddings of negative patterns)**.** *Let* $\boldsymbol{s} = \langle s_1 \ldots s_n \rangle$ *be a sequence and* $\boldsymbol{p} = \langle p_1\ \neg q_1\ \ldots\ \neg q_{m-1}\ p_m \rangle$ *be a negative sequential pattern.*

An increasing[4] *tuple* $\boldsymbol{e} = (e_i)_{i \in [m]} \in [n]^m$ *is a* $\circ$*-embedding (read:* ***soft-embedding****) of pattern* $\boldsymbol{p}$ *in sequence* $\boldsymbol{s}$ *iff:*

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

An increasing[4] *tuple* $\boldsymbol{e} = (e_i)_{i \in [m]} \in [n]^m$ *is a* $\bullet$*-embedding (read:* ***strict-embedding****) of pattern* $\boldsymbol{p}$ *in sequence* $\boldsymbol{s}$ *iff:*

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

Intuitively, the constraint of a negative itemset q_i is checked on the sequence's itemsets at positions in interval $[e_i + 1, e_{i+1} - 1]$, *i.e.* between occurrences of the two positive itemsets surrounding the negative itemset in the pattern. A soft embedding considers individually each of the sequence's itemsets of $[e_i + 1, e_{i+1} - 1]$ while a strict embedding consider them as a whole.

Example 2 (Itemset absence semantics)**.** *Let* $\boldsymbol{p} = \langle a\ \neg(bc)\ d \rangle$ *be a pattern and consider four sequences as follows:*

Sequence	$\not\subseteq_D$	$\not\subseteq_D$	$\not\subseteq_G$	$\not\subseteq_G$
	$\bullet$	$\circ$	$\bullet$	$\circ$
$\boldsymbol{s_1} = \langle a\ c\ b\ e\ d \rangle$				✓
$\boldsymbol{s_2} = \langle a\ (bc)\ e\ d \rangle$				
$\boldsymbol{s_3} = \langle a\ b\ e\ d \rangle$			✓	✓
$\boldsymbol{s_4} = \langle a\ e\ d \rangle$	✓	✓	✓	✓

The reader can notice that each sequence contains a unique occurrence of $\boldsymbol{p}^+ = \langle a\ d \rangle$*, the positive part of pattern* $\boldsymbol{p}$*. Considering soft-embedding and partial non-inclusion (*$\not\subseteq_* := \not\subseteq_G$*),* $\boldsymbol{p}$ *occurs in* $\boldsymbol{s_1}$*,* $\boldsymbol{s_3}$ *and* $\boldsymbol{s_4}$ *but not in* $\boldsymbol{s_2}$*. Considering strict-embedding and partial non-inclusion,* $\boldsymbol{p}$ *occurs in* $\boldsymbol{s_3}$ *and* $\boldsymbol{s_4}$*. Indeed, items b and c occur between occurrences of a and d in* $\boldsymbol{s_1}$ *and* $\boldsymbol{s_2}$*. Considering total non-inclusion (*$\not\subseteq_* := \not\subseteq_D$*) and either type of embeddings, the absence of an itemset is satisfied if any of its items is absent. Hence,* $\boldsymbol{p}$ *occurs only in* $\boldsymbol{s_4}$*.*

Lemma 2. *If* $\boldsymbol{e}$ *is a* $\bullet$*-embedding, then* $\boldsymbol{e}$ *is a* $\circ$*-embedding, regardless of whether* $\not\subseteq_*$ *is* $\not\subseteq_G$ *or* $\not\subseteq_D$*.*

Lemma 3. *In the case that* $\not\subseteq_*$ *is* $\not\subseteq_D$*,* $\boldsymbol{e}$ *is a* $\circ$*-embedding iff* $\boldsymbol{e}$ *is a* $\bullet$*-embedding*

Lemma 4. *Let* $p = \langle p_1\ \neg q_1\ \ldots\ \neg q_{n-1}\ p_n \rangle \in \mathcal{N}$ *such that* $|q_i| \leq 1$ *for all* $i \in [n-1]$*, then* $\boldsymbol{e}$ *is a* $\circ$*-embedding iff* $\boldsymbol{e}$ *is a* $\bullet$*-embedding.*

Lemma 4 shows that in the simple case of patterns where all negative itemsets are singleton sets, the notions of strict and soft embeddings coincide.

Lemma 5. *Let* $\boldsymbol{p} \in \mathcal{N}$*, if* $\boldsymbol{e}$ *is an embedding of* $\boldsymbol{p}$ *in some sequence* $\boldsymbol{s}$*, then* $\boldsymbol{e}$ *is an embedding of* $\boldsymbol{p}^+$ *in* $\boldsymbol{s}$*.*

Another point that determines the semantics of negative containment concerns the multiple occurrences of some pattern in a sequence: should at least one or should all occurrences of the pattern positive part in the sequence satisfy the non-inclusion constraints?

Definition 5 (Negative pattern occurrence)**.** *Let* $\boldsymbol{s}$ *be a sequence and* $\boldsymbol{p}$ *be a negative sequential pattern with* $\boldsymbol{p}^+$ *the positive part of* $\boldsymbol{p}$*. For* $\not\subseteq_* \in \{\not\subseteq_D, \not\subseteq_G\}$ *and* $\LEFTcircle \in \{\circ, \bullet\}$*,*

- $\boldsymbol{p} \preceq^*_{\LEFTcircle} \boldsymbol{s}$ *denotes that pattern* $\boldsymbol{p}$ *occurs in sequence* $\boldsymbol{s}$ *iff there exists at least one* $\LEFTcircle$*-embedding of* $\boldsymbol{p}$ *in* $\boldsymbol{s}$ *considering the* $\not\subseteq_*$ *non-inclusion.*
- $\boldsymbol{p} \sqsubseteq^*_{\LEFTcircle} \boldsymbol{s}$ *denotes that pattern* $\boldsymbol{p}$ *occurs in sequence* $\boldsymbol{s}$ *iff for each embedding* $\boldsymbol{e}$ *of* $\boldsymbol{p}^+$ *in* $\boldsymbol{s}$*,* $\boldsymbol{e}$ *is also a* $\LEFTcircle$*-embedding of* $\boldsymbol{p}$ *in* $\boldsymbol{s}$ *considering the* $\not\subseteq_*$ *non-inclusion, and there exists at least one embedding* $\boldsymbol{e}$ *of* $\boldsymbol{p}^+$*.*

Definition 5 permits to capture two semantics for negative sequential patterns depending on the occurrences of the positive part: $\boldsymbol{p} \sqsubseteq^*_{\LEFTcircle} \boldsymbol{s}$ states that a negative pattern $\boldsymbol{p}$ occurs in a sequence $\boldsymbol{s}$ iff there exists at least one occurrence of the positive part of pattern $\boldsymbol{p}$ in sequence $\boldsymbol{s}$ and **every** such occurrence satisfies the negative constraints; $\boldsymbol{p} \preceq^*_{\LEFTcircle} \boldsymbol{s}$ states that $\boldsymbol{p}$ occurs in a sequence $\boldsymbol{s}$ iff there exists at least one occurrence of the positive part of pattern $\boldsymbol{p}$ in sequence $\boldsymbol{s}$ and **at least one** of these occurrences satisfies the negative constraints.

Example 3 (Strong vs weak occurrence semantics)**.** *Let* $\boldsymbol{p} = \langle a\ b\ \neg c\ d \rangle$ *be a pattern,* $\boldsymbol{s_1} = \langle a\ b\ e\ d \rangle$ *and* $\boldsymbol{s_2} = \langle a\ b\ c\ a\ d\ e\ b\ d \rangle$ *be two sequences. Thus,* $\boldsymbol{p}^+ = \langle a\ b\ d \rangle$ *occurs once in* $\boldsymbol{s_1}$ *hence there is no difference for occurrences of* $\boldsymbol{p}$ *in* $\boldsymbol{s_1}$ *under the two semantics. However,* $\boldsymbol{p}^+$ *occurs four times in* $\boldsymbol{s_2}$ *through embeddings* $(1, 2, 5)$*,* $(1, 2, 8)$*,* $(1, 7, 8)$ *and* $(4, 7, 8)$*. The first two occurrences do not satisfy the negative constraint (*$\neg c$*) but the last two occurrences do. Under the weak occurrence semantics, pattern* $\boldsymbol{p}$ *occurs in sequence* $\boldsymbol{s_2}$ *whereas it fails to do so under the strong occurrence semantics.*

Lemma 6. *Let* $\boldsymbol{p}$ *be an NSP and* $\boldsymbol{s}$ *a sequence. For* $\LEFTcircle \in \{\circ, \bullet\}$ *and* $\not\subseteq_* \in \{\not\subseteq_D, \not\subseteq_G\}$*,*

$$\boldsymbol{p} \sqsubseteq^*_{\LEFTcircle} \boldsymbol{s} \implies \boldsymbol{p} \preceq^*_{\LEFTcircle} \boldsymbol{s} \tag{2}$$

Lemma 7. *Let* $\boldsymbol{p}$ *be an NSP and* $\boldsymbol{s}$ *a sequence. For* $\LEFTcircle \in \{\circ, \bullet\}$*,*

$$\boldsymbol{p} \preceq^D_{\LEFTcircle} \boldsymbol{s} \implies \boldsymbol{p} \preceq^G_{\LEFTcircle} \boldsymbol{s}$$
$$\boldsymbol{p} \sqsubseteq^D_{\LEFTcircle} \boldsymbol{s} \implies \boldsymbol{p} \sqsubseteq^G_{\LEFTcircle} \boldsymbol{s}$$

In this section, we have exhibited several semantics that can be associated to negative patterns. This leads to eight different types of pattern occurrences. We take Θ to denote the set of containment relations:

$$\Theta = \left\{ \preceq^D_\circ, \preceq^D_\bullet, \preceq^G_\circ, \preceq^G_\bullet, \sqsubseteq^D_\circ, \sqsubseteq^D_\bullet, \sqsubseteq^G_\circ, \sqsubseteq^G_\bullet \right\}$$

These containment relations allow to disambiguate the semantics of negative pattern containment encountered in the literature. Next, Section 4 investigates possible equivalent containment relations in Θ.

4 Dominance and equivalence between containment relations

Definition 6 (Dominance)**.** *For* $\theta, \theta' \in \Theta$*,* θ dominates θ'*, denoted* $\theta \geqslant \theta'$*, iff* $\boldsymbol{p}\theta\boldsymbol{s} \implies \boldsymbol{p}\theta'\boldsymbol{s}$ *for all* $\boldsymbol{p} \in \mathcal{N}$ *and all sequence* $\boldsymbol{s}$*.*

We denote by $\theta \not\geqslant \theta'$ *iff* $\theta \geqslant \theta'$ *is false, that is, there is some couple* $(\boldsymbol{p}, \boldsymbol{s})$ *such that* $\boldsymbol{p}\theta\boldsymbol{s}$ *but not* $\boldsymbol{p}\theta'\boldsymbol{s}$*.*

[4] By an increasing tuple $\boldsymbol{e}$, we mean a tuple such that $e_i < e_{i+1}$ (in particular, repetitions are not allowed).

The idea behind dominance between two containment relations θ and θ' is related to the sequences in which a pattern occurs. By definition, if $\theta \geqslant \theta'$ then a pattern $\boldsymbol{p} \in \mathcal{N}$ occurs in a sequence $\boldsymbol{s}$ according to the θ' containment relation whenever $\boldsymbol{p}$ occurs in $\boldsymbol{s}$ according to the θ containment relation. In the context of pattern mining, this is useful to design algorithms exploiting properties of a dominating containment relation in order to extract efficiently the patterns according to dominated containment relations.

Lemma 8. *The dominance relation $\geqslant$ is a pre-order.*

Definition 7 (Equivalent containment relations)**.** *For $\theta, \theta' \in \Theta$, θ is equivalent to θ', denoted $\theta \sim \theta'$ iff $\theta \geqslant \theta'$ and $\theta' \geqslant \theta$.*

Lemma 9. *$\sim$ is an equivalence relation on Θ.*

Two equivalent containment relations have equivalent semantics, in the following sense: the sets of sequences in which a given pattern occurs are the same and, reciprocally, the sets of negative patterns that occur in a sequence are the same when considering these two containment relations.

We now study the dominance relations that hold between the elements of Θ.

Proposition 1. *The following dominance statements between containment relations hold:*

$$\sqsubseteq^{*}_{\LEFTcircle} \geqslant \preceq^{*}_{\LEFTcircle} \tag{3}$$

$$\preceq^{*}_{\bullet} \geqslant \preceq^{*}_{\circ} \text{ and } \sqsubseteq^{*}_{\bullet} \geqslant \sqsubseteq^{*}_{\circ} \tag{4}$$

$$\preceq^{D}_{\circ} \geqslant \preceq^{D}_{\bullet} \text{ and } \sqsubseteq^{D}_{\circ} \geqslant \sqsubseteq^{D}_{\bullet} \tag{5}$$

$$\preceq^{D}_{\LEFTcircle} \geqslant \preceq^{G}_{\LEFTcircle} \text{ and } \sqsubseteq^{D}_{\LEFTcircle} \geqslant \sqsubseteq^{G}_{\LEFTcircle} \tag{6}$$

and the following non-dominance statements hold:

$$\preceq^{G}_{\LEFTcircle} \ngeqslant \preceq^{D}_{\LEFTcircle} \text{ and } \sqsubseteq^{G}_{\LEFTcircle} \ngeqslant \sqsubseteq^{D}_{\LEFTcircle} \tag{7}$$

$$\preceq^{*}_{\LEFTcircle} \ngeqslant \sqsubseteq^{*}_{\LEFTcircle} \tag{8}$$

$$\preceq^{G}_{\circ} \ngeqslant \preceq^{G}_{\bullet} \text{ and } \sqsubseteq^{G}_{\circ} \ngeqslant \sqsubseteq^{G}_{\bullet} \tag{9}$$

$$\preceq^{G}_{\bullet} \ngeqslant \sqsubseteq^{G}_{\circ} \tag{10}$$

$$\sqsubseteq^{G}_{\circ} \ngeqslant \preceq^{G}_{\bullet} \tag{11}$$

$$\preceq^{*}_{\bullet} \ngeqslant \sqsubseteq^{*'}_{\bullet} \tag{12}$$

$$\preceq^{D}_{\circ} \ngeqslant \sqsubseteq^{G}_{\circ} \tag{13}$$

$$\sqsubseteq^{G}_{\bullet} \ngeqslant \preceq^{D}_{\bullet} \tag{14}$$

where $\not\subseteq_* \in \left\{\not\subseteq_D, \not\subseteq_G\right\}$ and $\LEFTcircle \in \{\circ, \bullet\}$.

Proposition 1 gathers results from Section 3. Each line expresses several relationships between pairs of containment relations. Equations 4-6 are dominance statements deduced from Lemmas 2, 3, 6 and 7. Equations 7-9 state the absence of dominance for which we can exhibit counterexamples. In addition, many other dominance and non dominance relationships can be deduced from Proposition 1 using transitivity of dominance (Lemma 8). Table 2 summarizes them.

An interesting result in Proposition 1 is that there are two pairs of containment relations, $\left(\sqsubseteq^{D}_{\circ}, \sqsubseteq^{D}_{\bullet}\right)$ and $\left(\preceq^{D}_{\circ}, \preceq^{D}_{\bullet}\right)$, whose two members are equivalent. In fact, there are six equivalence classes of containment relations: $\left\{\sqsubseteq^{G}_{\circ}\right\}$, $\left\{\sqsubseteq^{G}_{\bullet}\right\}$, $\left\{\preceq^{G}_{\circ}\right\}$, $\left\{\preceq^{G}_{\bullet}\right\}$, $\left\{\sqsubseteq^{D}_{\circ}, \sqsubseteq^{D}_{\bullet}\right\}$ and $\left\{\preceq^{D}_{\circ}, \preceq^{D}_{\bullet}\right\}$. Figure 2 illustrates the dominance relation on the quotient set $\Theta/\!\sim$.

We can finally point out that Lemma 4 adds a dominance relationship for the case that, in negative sequential patterns, negative itemsets are restricted to be singleton sets. In this case, the equivalence classes become: $\left\{\preceq^{D}_{\circ}, \preceq^{D}_{\bullet}\right\}$, $\left\{\preceq^{G}_{\circ}, \preceq^{G}_{\bullet}\right\}$, $\left\{\sqsubseteq^{D}_{\circ}, \sqsubseteq^{D}_{\bullet}\right\}$ and $\left\{\sqsubseteq^{G}_{\circ}, \sqsubseteq^{G}_{\bullet}\right\}$. Figure 2 illustrates the dominance relation on the quotient set $\Theta/\!\sim$ in this specific case.

Figure 2. Dominance between containment relations. The labels for edges refer to the corresponding equations in Proposition 1. Dominance goes from top to bottom i.e. $\sqsubseteq^{D}_{\circ}$ as well as $\sqsubseteq^{D}_{\bullet}$ dominate all other containment relations.

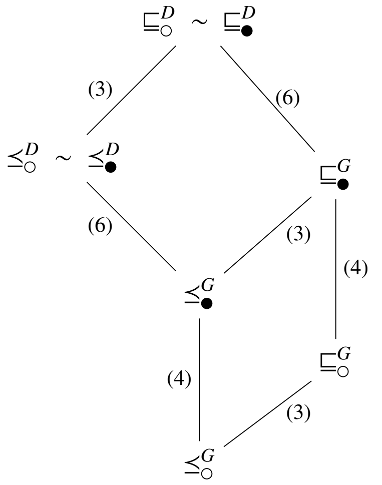
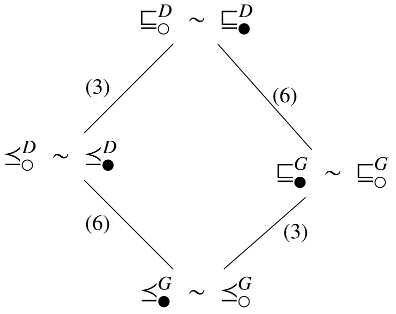

$\sqsubseteq^{D}_{\circ} \sim \sqsubseteq^{D}_{\bullet}$
(3)
(6)
$\preceq^{D}_{\circ} \sim \preceq^{D}_{\bullet}$
$\sqsubseteq^{G}_{\bullet}$
(6)
(3)
(4)
$\preceq^{G}_{\bullet}$
$\sqsubseteq^{G}_{\circ}$
(4)
(3)
$\preceq^{G}_{\circ}$

Figure 3. Dominance between containment relations for the case that, in negative sequential patterns, negative itemsets are restricted to be singleton sets. The labels for edges again refer to the equations given in Proposition 1.


$\sqsubseteq^{D}_{\circ} \sim \sqsubseteq^{D}_{\bullet}$
(3)
(6)
$\preceq^{D}_{\circ} \sim \preceq^{D}_{\bullet}$
$\sqsubseteq^{G}_{\bullet} \sim \sqsubseteq^{G}_{\circ}$
(6)
(3)
$\preceq^{G}_{\bullet} \sim \preceq^{G}_{\circ}$

5 Anti-monotonicity

It is now time to check whether there are containment relations that enjoy interesting properties. In our initial context of mining frequent negative sequential patterns, we investigate anti-monotonicity properties.

According to Wang et *al.* [12], "the downward property (. . .) does not hold in negative sequential analysis". The "downward property" denotes the anti-monotonicity property. We will see that this assertion is actually false with some semantics.

Anti-monotonicity makes sense only with a partial order on the set of NSPs. We first introduce different possible partial orders and then we introduce anti-monotonicity.

In the remaining of the section, non-inclusion of itemsets is total non-inclusion, $\not\subseteq_D$. Thus, we can count on the anti-monotonicity of non-inclusion of itemsets: $q \subseteq q' \implies \forall p \in \mathcal{I},\ (q' \not\subseteq_D p \Rightarrow q \not\subseteq_D p)$.

Table 2. Dominance. $\geqslant$ (resp. $-$) means that the semantics at the left of the row dominates (resp. does not dominate) the semantics at the top of the column.

	$\sqsubseteq^G_\bullet$	$\preceq^G_\bullet$	$\sqsubseteq^G_\circ$	$\preceq^G_\circ$	$\sqsubseteq^D_\bullet$	$\preceq^D_\bullet$	$\sqsubseteq^D_\circ$	$\preceq^D_\circ$
$\sqsubseteq^G_\bullet$	$\cdot$	$\geqslant$	$\geqslant$	$\geqslant$	$-$	$-$	$-$	$-$
$\preceq^G_\bullet$	$-$	$\cdot$	$-$	$\geqslant$	$-$	$-$	$-$	$-$
$\sqsubseteq^G_\circ$	$-$	$-$	$\cdot$	$\geqslant$	$-$	$-$	$-$	$-$
$\preceq^G_\circ$	$-$	$-$	$-$	$\cdot$	$-$	$-$	$-$	$-$
$\sqsubseteq^D_\bullet$	$\geqslant$	$\geqslant$	$\geqslant$	$\geqslant$	$\cdot$	$\geqslant$	$\geqslant$	$\geqslant$
$\preceq^D_\bullet$	$-$	$\geqslant$	$-$	$\geqslant$	$-$	$\cdot$	$-$	$\geqslant$
$\sqsubseteq^D_\circ$	$\geqslant$	$\geqslant$	$\geqslant$	$\geqslant$	$\geqslant$	$\geqslant$	$\cdot$	$\geqslant$
$\preceq^D_\circ$	$-$	$\geqslant$	$-$	$\geqslant$	$-$	$\geqslant$	$-$	$\cdot$

5.1 Partial orders

Definition 8 introduces three relations between negative sequential patterns that are partial orders (see Proposition 2).

Definition 8 (NSP relations)**.** *Consider two NSPs* $\boldsymbol{p} = \langle p_1\ \neg q_1\ p_2\ \neg q_2\ \cdots\ p_{k-1}\ \neg q_{k-1}\ p_k\rangle$ *and* $\boldsymbol{p'} = \langle p'_1\ \neg q'_1\ p'_2\ \neg q'_2\ \cdots\ p'_{k'-1}\ \neg q'_{k'-1}\ p'_{k'}\rangle$*.*

By definition, $\boldsymbol{p} \mathrel{\dot{\lhd}} \boldsymbol{p'}$ *iff* $k \leq k'$ *and there exists an increasing*[4] *tuple* $(u_i)_{i\in[k]} \in [k']^k$ *and:*

1. $\forall i \in [k],\ p_i \subseteq p'_{u_i}$
2. $\forall i \in [k-1],\ q_i \subseteq \bigcup_{j\in[u_i,u_{i+1}-1]} q'_j$
3. $k = k' \implies \exists j \in [k],\ p_j \neq p'_j$ *or* $\exists j \in [k-1],\ q_j \neq q'_j$

by definition, $\boldsymbol{p} \lhd \boldsymbol{p'}$ *iff* $k \leq k'$ *and:*

1. $\forall i \in [k],\ p_i \subseteq p'_i$
2. $\forall i \in [k-1],\ q_i \subseteq q'_i$
3. $k = k' \implies p_k \neq p'_k$ *or* $\exists j \in [k-1]$ *s.t.* $q_j \neq q'_j$

and, by definition, $\boldsymbol{p} \mathrel{\lhd\!\!\!+} \boldsymbol{p'}$ *iff* $k = k'$ *and:*

1. $\forall i \in [k],\ p_i = p'_i$
2. $\forall i \in [k-1],\ q_i \subseteq q'_i$
3. $\exists j \in [k-1]$ *s.t.* $q_j \neq q'_j$

The $\mathrel{\dot{\lhd}}$ relation can be seen as the "classical" inclusion relation between sequential patterns [11]. An NSP $\boldsymbol{p}$ is less specific than $\boldsymbol{p'}$ iff $\boldsymbol{p}^+$ is a subsequence of $\boldsymbol{p'}^+$ and negative constraints are satisfied. The main difference with $\lhd$ is that $\mathrel{\dot{\lhd}}$ permits to insert new positive itemsets in the middle of the sequence while $\lhd$ permits only insertion of new positive itemsets at the end.[5,6] Nonetheless, it is still possible to insert items to the positive itemsets. The $\mathrel{\lhd\!\!\!+}$ does not even permit such differences: for two NSPs to be comparable via $\mathrel{\lhd\!\!\!+}$, they must have the same positive itemsets.

Lemma 10. *For* $\boldsymbol{p}, \boldsymbol{p'} \in \mathcal{N}$,

$$\boldsymbol{p} \mathrel{\lhd\!\!\!+} \boldsymbol{p'} \implies \boldsymbol{p} \lhd \boldsymbol{p'} \implies \boldsymbol{p} \mathrel{\dot{\lhd}} \boldsymbol{p'} \tag{15}$$

Proposition 2 (Strict partial orders)**.** $\mathrel{\dot{\lhd}}$*,* $\lhd$ *and* $\mathrel{\lhd\!\!\!+}$ *are partial orders on* $\mathcal{N}$*.*

We can notice that the third conditions in Definition 8 enforce the relations to be irreflexive. Removing these conditions enables to define non-strict partial orders.

5.2 Anti-monotonicity

Let us first define the anti-monotonicity property of a containment relation $\theta \in \Theta$ considering a strict partial order $\ltimes \in \{\lhd, \mathrel{\dot{\lhd}}, \mathrel{\lhd\!\!\!+}\}$.

Definition 9 (Anti-monotonicity on $(\mathcal{N}, \ltimes)$)**.** *Let* $\theta \in \Theta$ *be a containment relation,* θ *is anti-monotonic on* $(\mathcal{N}, \ltimes)$ *iff for all* $\boldsymbol{p}, \boldsymbol{p'} \in \mathcal{N}$ *and all sequences* $\boldsymbol{s}$*:*

$$\boldsymbol{p} \ltimes \boldsymbol{p'} \implies (\boldsymbol{p'}\theta\boldsymbol{s} \implies \boldsymbol{p}\theta\boldsymbol{s})$$

First of all, we provide an example showing that none of the containment relations is anti-monotonic on $(\mathcal{N}, \mathrel{\dot{\lhd}})$. Let $\boldsymbol{p} = \langle b\ \neg c\ a\rangle$, $\boldsymbol{p'} = \langle b\ \neg c\ d\ a\rangle$ and $\boldsymbol{s} = \langle b\ e\ d\ c\ a\rangle$. Then, we have $\boldsymbol{p} \mathrel{\dot{\lhd}} \boldsymbol{p'}$.[7] Nonetheless, for each $\theta \in \Theta$, $\boldsymbol{p'}\theta\boldsymbol{s}$ but it is false that $\boldsymbol{p}\theta\boldsymbol{s}$.

In fact, the presence of the item d in the sequence changes the scope for checking the absence of c. This example is similar to the one used by Zheng et *al.* [15] to state that anti-monotonic property does not hold for negative sequential patterns. Nonetheless, the anti-monotonicity property holds in case the partial order prevents from changing the scope for absent items.

Proposition 3. $\preceq^D_\circ$ *and* $\preceq^D_\bullet$ *are anti-monotonic on* $(\mathcal{N}, \lhd)$*.*

Proposition 3 shows that using the $\lhd$ partial order causes anti-monotonicity to hold for containment with weak-occurrence. It is not the case with strong-occurrence, though. Let us give a counterexample illustrating what can happen with strong-occurrence. Let $\boldsymbol{p} = \langle a\ \neg b\ c\rangle$, $\boldsymbol{p'} = \langle a\ \neg b\ c\ d\rangle$ and $\boldsymbol{s} = \langle a\ c\ d\ a\ b\ c\rangle$. Then, we have $\boldsymbol{p} \lhd \boldsymbol{p'}$.[8] Nonetheless, $\boldsymbol{p'} \sqsubseteq^D_{\LEFTcircle} \boldsymbol{s}$ holds but it is false that $\boldsymbol{p} \sqsubseteq^D_{\LEFTcircle} \boldsymbol{s}$. In fact, without the presence of the item d in the pattern, there are three possible embeddings of $\boldsymbol{p}$ in $\boldsymbol{s}$. For $\sqsubseteq^D_{\LEFTcircle}$ each embedding must satisfy the negation of b, which is not the case, but for $\preceq^D_{\LEFTcircle}$ it is sufficient to have only one embedding satisfying negations.

The previous example illustrates the problem when extending the pattern with additional itemsets. The same issue is encountered with the following example considering patterns of equal length while one pattern has an extended itemset. Let $\boldsymbol{p} = \langle a\ \neg b\ c\rangle$, $\boldsymbol{p'} = \langle a\ \neg b\ (cd)\rangle$ and $\boldsymbol{s} = \langle a\ (cd)\ a\ b\ c\rangle$. Then, we have $\boldsymbol{p} \lhd \boldsymbol{p'}$. Nonetheless, $\boldsymbol{p'} \sqsubseteq^D_{\LEFTcircle} \boldsymbol{s}$ holds but it is false that $\boldsymbol{p} \sqsubseteq^D_{\LEFTcircle} \boldsymbol{s}$.

Proposition 4. $\preceq^D_\circ$, $\preceq^D_\bullet$, $\sqsubseteq^D_\circ$ *and* $\sqsubseteq^D_\bullet$ *are anti-monotonic on* $(\mathcal{N}, \mathrel{\lhd\!\!\!+})$*.*

We remind that this section was restricted to the case of total non-inclusion ($\not\subseteq_D$) but the results also hold when $\not\subseteq_D$ is replaced by $\not\subseteq_G$ except that we must reverse the inclusion relations for negatives in the partial orders (that is, $\bigcup_{j\in[u_i,u_{i+1}-1]} q'_j \subset q_i$ for $\mathrel{\dot{\lhd}}$ and $q'_i \subseteq q_i$ for $\lhd$ and $\mathrel{\lhd\!\!\!+}$).

[5] In sequential pattern mining, it is called a *backward*-extension of the patterns.

[6] We remind that, by Definition 1, $p_i \neq \emptyset$ and that we never have two successive negative itemsets in an NSP.

[7] In this case, we do not have $\boldsymbol{p} \lhd \boldsymbol{p'}$ nor $\boldsymbol{p} \mathrel{\lhd\!\!\!+} \boldsymbol{p'}$

[8] In this case, we also have $\boldsymbol{p} \mathrel{\dot{\lhd}} \boldsymbol{p'}$ (see Lemma 10) but not $\boldsymbol{p} \mathrel{\lhd\!\!\!+} \boldsymbol{p'}$

6 Application to pattern mining

The definitions of pattern support, frequent pattern and pattern mining derive naturally from the notion of occurrence of a negative sequential pattern, no matter the choices for embedding (soft or strict), non-inclusion (partial or total) and occurrences (weak or strong). However, these choices about the semantics of NSPs impact directly the number of frequent patterns (under the same minimal threshold constraint) and also computation time. The stronger the negative constraints, the fewer the number of sequences containing a pattern, and the lesser the number of frequent patterns.

Definition 10 (Pattern supports). *Let $\mathcal{D} = \{s_i\}_{i\in[n]}$ be a dataset of sequences and* $\boldsymbol{p}$ *be an NSP. The support of* $\boldsymbol{p}$ *in* $\mathcal{D}$*, denoted* $supp\text{–}\theta^{\mathcal{D}}(p)$*, is the number of sequences of* $\mathcal{D}$ *in which* $\boldsymbol{p}$ *occurs according to the* $\theta \in \Theta$ *containment relation.*

When there is no ambiguity on the dataset of sequences, $supp\text{–}\theta^{\mathcal{D}}(p)$ is denoted $supp\text{–}\theta(p)$.

Clearly, if a containment relation θ is dominated by another containment relation θ', then the support of the pattern evaluated with θ is lower than the support of the pattern evaluated with θ'. The next proposition ensues from Proposition 1.

Proposition 5. *For* $\boldsymbol{p} \in \mathcal{N}$*,*

$$supp-\sqsubseteq^{*}_{\LEFTcircle}(\boldsymbol{p}) \leq supp-\preceq^{*}_{\LEFTcircle}(\boldsymbol{p}) \tag{16}$$

$$\begin{aligned} supp-\sqsubseteq^{*}_{\bullet}(\boldsymbol{p}) &\leq supp-\sqsubseteq^{*}_{\circ}(\boldsymbol{p}) \\ supp-\preceq^{*}_{\bullet}(\boldsymbol{p}) &\leq supp-\preceq^{*}_{\circ}(\boldsymbol{p}) \end{aligned} \tag{17}$$

$$\begin{aligned} supp-\sqsubseteq^{D}_{\circ}(\boldsymbol{p}) &\leq supp-\sqsubseteq^{D}_{\bullet}(\boldsymbol{p}) \\ supp-\preceq^{D}_{\circ}(\boldsymbol{p}) &\leq supp-\preceq^{D}_{\bullet}(\boldsymbol{p}) \end{aligned} \tag{18}$$

$$\begin{aligned} supp-\sqsubseteq^{D}_{\LEFTcircle}(\boldsymbol{p}) &\leq supp-\sqsubseteq^{G}_{\LEFTcircle}(\boldsymbol{p}) \\ supp-\preceq^{D}_{\LEFTcircle}(\boldsymbol{p}) &\leq supp-\preceq^{G}_{\LEFTcircle}(\boldsymbol{p}) \end{aligned} \tag{19}$$

In addition, the following anti-monotonicity properties of support measures ensue from Propositions 3 and 4.

Proposition 6. *For* $\boldsymbol{p}, \boldsymbol{p}' \in \mathcal{N}$*,*

$$\boldsymbol{p} \lhd \boldsymbol{p}' \implies supp-\preceq^{D}_{\LEFTcircle}(\boldsymbol{p}') \leq supp-\preceq^{D}_{\LEFTcircle}(\boldsymbol{p}) \tag{20}$$

$$\boldsymbol{p} \mathrel{\lhd\!\!\!\!{\scriptscriptstyle +}} \boldsymbol{p}' \implies \begin{cases} supp-\sqsubseteq^{D}_{\LEFTcircle}(\boldsymbol{p}') \leq supp-\sqsubseteq^{D}_{\LEFTcircle}(\boldsymbol{p}) \\ supp-\preceq^{D}_{\LEFTcircle}(\boldsymbol{p}') \leq supp-\preceq^{D}_{\LEFTcircle}(\boldsymbol{p}) \end{cases} \tag{21}$$

There are two practical ways to exploit these results to implement efficient frequent NSP mining algorithms. On the one hand, the results from Proposition 6 can be directly used to implement algorithms with efficient and correct strategies to prune the search space.[9] For $\preceq^{D}_{\LEFTcircle}$ containment relation, Equation 20 exploits the $\lhd$ partial order to early prune a priori unfrequent patterns. For $\sqsubseteq^{D}_{\LEFTcircle}$ containment relation, the $\mathrel{\lhd\!\!\!\!{\scriptscriptstyle +}}$ partial order must be used to ensure the correctness of the algorithm (Equation 21). Unfortunately, $\mathrel{\lhd\!\!\!\!{\scriptscriptstyle +}}$ is less interesting than $\lhd$ because there are fewer pairs of comparable patterns. On the other hand, the support evaluated with $\preceq^{D}_{\LEFTcircle}$ is an upper bound for the support of $\sqsubseteq^{D}_{\LEFTcircle}$ (Equation 16). Thus, it is possible also to prune patterns accessible with the partial order $\lhd$ without losing the correctness of the pruning strategy.

7 A proposal to disambiguate syntax of negative sequential patterns

The ¬ symbol is overloaded in the literature about negative sequential pattern mining. Our intuition was that the different approaches [3, 7, 8, 9, 15] do not extract the same set of patterns because of slightly different definition of negative patterns. Our framework deals with the need to define an unambiguous containment relation between a negative sequential pattern $\boldsymbol{p}$ and a sequence $\boldsymbol{s}$ that informs the user about:

- how multiple occurrences of the positive part of $\boldsymbol{p}$ are handled,
- how negative itemsets are handled (type of embedding and type of non-inclusion relation).

We separate these two dimensions of our definition of a containment relation because the second refers to single itemsets, while the first refers to the whole pattern.

Thanks to our framework, we are able to assign a containment relation to each approach from the literature. The approaches based on eNSP [3, 7] are based on the containment relation $\sqsubseteq^{D}_{\bullet}$, PNSP [9] uses the relation $\preceq^{G}_{\bullet}$, and NegPSpan [8] and NegGSP [15] deal with the equivalent relations $\preceq^{D}_{\bullet}$ and $\preceq^{D}_{\circ}$.

This confirms our initial intuition: the different approaches do not use the same containment relations and thus they do not aim at extracting the same set of patterns. Moreover, it is worth noticing that these four approaches explore a large range of the possible containment relations. eNSP exploits the strong notion of occurrence while the other approaches exploit the weak notion. All but PNSP approaches are based on total non-inclusion. Strict-embedding ($\bullet$) is generally preferred to soft-embedding ($\circ$). eNSP ($\sqsubseteq^{D}_{\bullet}$) made the most restrictive choice by using the containment relation that dominates all the others. On the opposite, the two least restrictive choices ($\preceq^{G}_{\circ}$ and $\sqsubseteq^{G}_{\circ}$) have not been explored, presumably due to their obvious lack of suitable properties for pattern mining.

Finally, it is worth comparing negative sequential patterns with some formulas in Linear Temporal Logic on finite traces (LTLf) [6]. The question is to find specific LTLf formulas capturing our containment relations between any two patterns $\boldsymbol{p}$ and $\boldsymbol{s}$. Then, it is interesting to notice that containment relations based on soft-embedding have simple counterparts in the language of LTLf. Indeed, the soft-embedding constraint imposes each successive itemset of a sequence to not contain some negated items. The strict-embedding constraint, which requires to evaluate a union of items, does not fit well to the linearity of LTLf formulas.

8 Conclusions and perspectives

In this article, we investigated formal properties of the semantics of negation in sequential patterns to answer our two main questions.

1. *What is a proper support measure for negative sequential patterns?* We gave eight possible semantics and as many support measures. We can conclude that there is not a single way to evaluate the support of an NSP.
2. *Is there a support measure enjoying anti-monotonicity?* We run counter the state of the art by proposing three partial orders for which anti-monotonicity holds although only for some semantics of negative sequential pattern mining.

The combination of partial order $\lhd$ and containment relation $\preceq^{D}_{\LEFTcircle}$ appears to be a good candidate for developping a complete, correct

[9] Completeness and non-redundancy of algorithms are out of the scope of this article.

and non-redundant negative sequential pattern mining algorithm [8]. One advantage of an approach based on an anti-monotonic support measure is the benefit from decades of research in pattern mining so as to extend the mining of NSP to the mining of closed NSP or the mining of NSP with *maxgap* or *maxspan* constraints.

Nonetheless, no semantics is "more correct" or relevant than another one. It depends on the notion to be captured. Our objective is to give the opportunity to make an educated choice. It is especially important with NSP as the choice of a mining algorithm is not only a matter of computational efficiency, but also a matter of semantics.

In view of Definition 4 and Lemma 3, three possibilities arise for evaluating a negative itemset (syntactically distinguished below by writing a negative itemset $\neg(a_1, \dots, a_{l_i})$ or $\neg\{a_1, \dots, a_{l_i}\}$ or $\neg|a_1, \dots, a_{l_i}|$) as follows:

$\neg(a_1, \dots, a_{l_i})$ is evaluated as

$$\{a_1, \dots, a_{l_i}\} \nsubseteq_G s_j,\ \forall j \in [e_i + 1, e_{i+1} - 1] \text{ for all } i \in [m-1]$$

Intuitively, you check that, in between s_{e_i} (i.e., a match for p_i) and $s_{e_{i+1}}$ (i.e., a match for p_{i+1}), none of these s_j include all of $a_1, \dots, a_{l_i}$.

$\neg\{a_1, \dots, a_{l_i}\}$ is evaluated as

$$\{a_1, \dots, a_{l_i}\} \nsubseteq_G \bigcup_{j \in [e_i+1, e_{i+1}-1]} s_j \text{ for all } i \in [m-1]$$

Intuitively, you check that there exists some item in $a_1, \dots, a_{l_i}$ that does not occur at all in between s_{e_i} (i.e., a match for p_i) and $s_{e_{i+1}}$ (i.e., a match for p_{i+1}).

$\neg|a_1, \dots, a_{l_i}|$ is evaluated as

$$\{a_1, \dots, a_{l_i}\} \nsubseteq_D \bigcup_{j \in [e_i+1, e_{i+1}-1]} s_j \text{ for all } i \in [m-1]$$

Intuitively, you check that every item in $a_1, \dots, a_{l_i}$ fails to occur in between s_{e_i} (*i.e.* , a match for p_i) and $s_{e_{i+1}}$ (i.e., a match for p_{i+1}).

This opens the way for a syntax of negative sequential patterns that is even more expressive. Indeed, it enables to mix different types of negation within a pattern. For instance, we can specify patterns such as $\langle a\ \neg|bc|\ f\ \neg\{ac\}\ b\rangle$ (intuitively: none of b and c occur between a and f; also, either a or c (or both) does not occur at all between f and b).

The first perspective of this work is to evaluate the proposed notations on a panel of real users. A preliminary survey concluded on a lack of any dominant interpretation of the ¬ symbol. We would like to confirm this preliminary result on a larger panel and to evaluate the benefit of having a dedicated syntax for each containment relation.

Our second perspective is to extend our theoretical results from the pattern recognition perspective. Matching sequential patterns in sequences is a fundamental issue in monitoring of discrete event systems, in genetic data analysis, in text analysis, etc. Adding negations to sequential patterns increases the expressivity of the pattern language. It raises questions about space and time complexity of the recognition and/or enumeration of negative sequential patterns: are the different containment relations equally hard to evaluate in sequences?

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant, 'Fast Algorithms for Mining Association Rules in Large Databases', in *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pp. 487–499, (1994).

[2] Pedro Cabalar, David Pearce, and Agustín Valverde, 'Stable reasoning', *Journal of Applied Non-Classical Logics*, **27**(3-4), 238–254, (2017).

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

[4] Longbing Cao, Philip S. Yu, and Vipin Kumar, 'Nonoccurring behavior analytics: A new area', *Intelligent Systems*, **30**(6), 4–11, (2015).

[5] Yen-Liang Chen, Mei-Ching Chiang, and Ming-Tat Ko, 'Discovering time-interval sequential patterns in sequence databases', *Expert System with Applications*, **25**(3), 343–354, (2003).

[6] Giuseppe De Giacomo and Moshe Y Vardi, 'Linear temporal logic and linear dynamic logic on finite traces', in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, (2013).

[7] Yongshun Gong, Tiantian Xu, Xiangjun Dong, and Guohua Lv, 'e-NSPFI: Efficient mining negative sequential pattern from both frequent and infrequent positive sequential patterns', *International Journal of Pattern Recognition and Artificial Intelligence*, **31**(02), 1750002, (2017).

[8] 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).

[9] Sue-Chen Hsueh, Ming-Yen Lin, and Chien-Liang Chen, 'Mining negative sequential patterns for e-commerce recommendations', in *Proceedings of Asia-Pacific Services Computing Conference*, pp. 1213–1218, (2008).

[10] Tomasz Imielinski and Heikki Mannila, 'A database perspective on knowledge discovery', *Communications of the ACM*, **39**(11), 58–64, (1996).

[11] Carl H. Mooney and John F. Roddick, 'Sequential pattern mining – approaches and algorithms', *ACM Computing Survey*, **45**(2), 1–39, (2013).

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

[13] Heinrich Wansing, *Negation*, chapter 18, 415–436, John Wiley & Sons, Ltd, 2017.

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

[15] Zhigang Zheng, Yanchang Zhao, Ziye Zuo, and Longbing Cao, 'Negative-GSP: An efficient method for mining negative sequential patterns', in *Proceedings of the Australasian Data Mining Conference*, pp. 63–67, (2009).