Declarative mining of negative sequential patterns

Philippe Besnard, Thomas Guyet

To cite this version:

Philippe Besnard, Thomas Guyet. Declarative mining of negative sequential patterns. 1st Declarative Problem Solving Workshop (DPSW 2020) @ ECAI 2020, Pedro Cabalar (CITIC-Univ. A Coruña, Spain); Andreas Herzig (Tolouse Univ, France); David Pearce (Univ. Politécnica Madrid, Spain); Torsten Schaub (Univ. Potsdam, Germany); Stefan Woltran (Vienna Univ. Tech. Austria), Aug 2020, Santiago de Compostela, Spain. pp.1-8. hal-03025560

HAL Id: hal-03025560
https://inria.hal.science/hal-03025560
Submitted on 26 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Declarative mining of negative sequential patterns

Philippe Besnard 1 and Thomas Guyet 2

Abstract. Declarative pattern mining consists in using declarative frameworks to solve pattern mining tasks. In this article, we address the task of mining negative sequential patterns in Answer Set Programming (ASP). 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 \( (a \rightarrow b \leftarrow c) \) arises with an occurrence of \( a \) and a subsequent occurrence of \( c \) but no occurrence of \( b \) in between. Recent results shed light on the ambiguity of such a seemingly intuitive notation, exhibited three semantics of the negative events and proposed alternative notations for them. In this article, we propose Answer Set Programming encodings of these three semantics in order to extract frequent negative sequential patterns from a set of sequences. It relies on previous encodings of frequent sequential pattern mining. We experiment with our encoding on synthetic data and compare the numbers of extracted patterns and the computing time obtained for each kind of negation. Surprisingly, the semantics that has the best algorithmic properties for pattern mining is not associated to an encoding that is the most efficient.

1 Introduction

Pattern mining [11] is a data mining task that consists in extracting interesting structured patterns from a set of structured examples. The standard pattern mining tasks are itemset mining, sequence mining and graph mining. The interestingness measure of a pattern is, in most of the algorithms, the number of its occurrences in the set of examples. Given a threshold \( k \), interesting patterns are those that occur at least in \( k \) examples. In this case, the task is known as frequent pattern mining for which many algorithms have been proposed. Most of the efficient algorithmic solutions rely on an antimoniomonicity property of the support: the larger the pattern, the fewer it occurs.

Declarative pattern mining aims at encoding pattern mining tasks in a declarative framework, and more specifically the frequent pattern mining tasks. Declarative pattern mining addressed the tasks of frequent itemset mining [5, 13, 14], frequent sequential patterns [15, 2, 1] or frequent graph mining [6]. Different declarative frameworks have been explored: Constraint Programming (CP) [5, 15, 6], SAT [12, 1] and Answer Set Programming (ASP) [13, 2]. Declarative pattern mining does not expect to be competitive with dedicated algorithms, but to take advantage of the versatility of declarative frameworks to propose pattern mining tools that could exploit background knowledge during the mining process to extract less but meaningful patterns.

In this work, we explore negative sequential patterns [8, 10, 17]. Standard sequential pattern mining algorithms [14] extract sequential patterns that frequently occur in a set of sequences. A sequential pattern is a sequence of events or items. For example, the sequential pattern \( (a \leftarrow c \rightarrow d) \) is read as “\( a \) occurs and then \( c \) occurs and finally \( d \) occurs”. Negative sequential patterns are sequential patterns that also specify non-occurring events. Intuitively, the syntax of a simple negative sequential pattern is as follows: \( (a \rightarrow b \leftarrow c) \). 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. Guyet and Besnard [8] highlighted that \( \neg \) symbol has several semantics but only few of them have good properties for pattern mining.

In this article, we propose an ASP encoding of the mining of frequent negative sequential patterns. Our encoding relies on previous work on frequent sequential pattern mining in ASP [9]. The objective of this article is also to illustrate the semantics of negative sequential patterns through declarative encodings.

2 Background

2.1 Sequential pattern mining

Our terminology on sequence mining follows the one in [15]. Throughout this article, \( \{1, \ldots, n\} \) denotes the set of the first \( n \) positive integers.

Let \( \Sigma \) be the set of items (alphabet). An itemset \( A = \{a_1, a_2, \ldots, a_m\} \subseteq \Sigma \) is a finite set of items. The size of \( A \), denoted \( |A| \), is \( m \). A sequence \( s \) is of the form \( s = (s_1, s_2, \ldots, s_n) \) where each \( s_i \) is an itemset. \( n \) is the length of the sequence.

A database \( D \) is a multi-set of sequences over \( \Sigma \). A sequence \( s = (s_1, s_2, \ldots, s_m) \) with \( m \leq n \), written \( s \subseteq t \), is contained in a sequence \( t = (t_1, t_2, \ldots, t_n) \) with \( n \leq m \) for \( 1 \leq i \leq m \) and an increasing sequence \((e_1, e_2, \ldots, e_m)\) of positive integers \( e_i \in \{n\} \), called an embedding of \( s \) in \( t \). For example, we have \( (a (c d)) \subseteq (a b (c d e)) \). Relative to embedding \((1, 3)\), \((c d)\) denotes the itemset made of items \( c \) and \( d \).

Given a database \( D \), the cover of a sequence \( p \) is the set of sequences in \( D \) that contain \( p \): \( \text{cover}(p, D) = \{ t \in D | p \subseteq t \} \). The number of sequences in \( D \) containing \( p \) is called its support, that is, \( \text{support}(p, D) = |\text{cover}(p, D)| \). For an integer \( k \), frequent sequence mining is about discovering all sequences \( p \) such that \( \text{support}(p, D) \geq k \). We often call \( p \) a (sequential) pattern, and \( k \) is also referred to as the (minimum) frequency threshold.

2.2 Answer set programming

A logic program is a set of rules of the form

\[ a_0 : - a_1; \ldots ; a_k; \not a_{k+1}; \ldots ; \not a_n. \]  

where each \( a_i \) is a propositional atom for \( 0 \leq i \leq n \) and \( \not \) stands for default negation. If \( n = 0 \), rule (1) is called a fact. If \( a_0 \) is omitted, (1) represents an integrity constraint. Semantically, a

Listing 1. Facts specifying a database of sequences

logic program induces a collection of so-called answer sets, which are distinguished models of the program determined by answer sets semantics; see [4] for details. To facilitate the use of ASP in practice, several extensions have been developed. First of all, rules with variables are viewed as shorthands for the set of their ground instances. Further language constructs include conditional literals and cardinality constraints [15]. The former are of the form a : b for which the corresponding instance of b(x) holds. Similarly, 2 { a(X) : b(x) } 4 holds whenever between two and four instances of a(X) (subject to b(x)) are true. Specifically, we rely on the sequel in the input language of the ASP system clingo [3].

3 ASP-based Sequence Mining

In this section, we introduce the sequential pattern mining tasks in ASP. The following programs correspond to the notion of skip-gaps encodings in [2,7]. Gebser et al. [3] uses a different encoding strategy for frequent sequential pattern that is not suitable for negative patterns.

3.1 Fact format

We represent a database D in terms of facts seq(t,p,e), saying that item e occurs at position p in a sequence t. For instance, Listing 1 specifies a database of seven sequences:

<table>
<thead>
<tr>
<th>Term</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq.</td>
<td>(d (ab) c)</td>
<td>(ac b c)</td>
<td>(a c b a)</td>
<td>((ab) c)</td>
<td>(a c)</td>
<td>(b c)</td>
<td>(c)</td>
</tr>
</tbody>
</table>

Considering the minimal frequency threshold \( k = 2 \), one can check that \( \langle a \rangle, \langle b \rangle, \langle c \rangle, \langle ab \rangle, \langle ac \rangle, \langle bc \rangle \) and \( \langle ab \rangle, \langle ac \rangle, \langle bc \rangle \) are the frequent patterns of the database.

3.2 Mining frequent sequences: basic encoding

The encoding principle for frequent sequence mining follows the one of [13], that is, each answer set comprises a single pattern of interest. In contrast to itemset mining, however, we need to take the order of items in a pattern into account to determine its support and check its frequency.

Listing 2 provides our basic encoding of frequent sequence mining borrowed from [2]. It relies on two parameters: \( \text{max} \) determines a maximum length for patterns of interest, and \( k \) specifies the minimal frequency threshold. An answer set then represents a frequent pattern \( p = (p_1 \ldots p_m) \) such that \( m \in [\text{max}] \) by atoms \( \text{pat}(m,p_1), \ldots, \text{pat}(m,p_m) \).

1 item(I) :- seq(T,P,I).
2 slot(I+1) :- slot(I); X=max.
3 \{ pat(X,E) : item(E) \} :- slot(X).
4 patlen(L) :- pat(len,E); not pat(len,E):item(E).
5 \{ patlen(L) \} :- pat(len,E); not pat(len,E):item(E).
6 emb(T,P,1) :- seq(T,P,E):pat(1,E); seq(T,P,_).
7 \{ emb(T,P,1) \} :- emb(T,Q,X); seq(T,P,_) \{ Q<P \}.
8 patlen(L) :- pat(len,E); not pat(len,E):item(E).
9 \{ emb(T,P,1) \} :- emb(T,Q,X); seq(T,P,_) \{ Q<P \}.
10 \{ emb(T,P,1) \} :- emb(T,Q,X); seq(T,P,_) \{ Q<P \}.
11 \{ emb(T,P,1) \} :- emb(T,Q,X); seq(T,P,_) \{ Q<P \}.
12 \{ emb(T,P,1) \} :- emb(T,Q,X); seq(T,P,_) \{ Q<P \}.
13 \{ emb(T,P,1) \} :- emb(T,Q,X); seq(T,P,_) \{ Q<P \}.
14 \{ emb(T,P,1) \} :- emb(T,Q,X); seq(T,P,_) \{ Q<P \}.
15 \{ emb(T,P,1) \} :- emb(T,Q,X); seq(T,P,_) \{ Q<P \}.
16 \{ emb(T,P,1) \} :- emb(T,Q,X); seq(T,P,_) \{ Q<P \}.

Listing 2. Basic encoding of frequent sequence mining

...,pat(1,p_m). That is, the first argument expresses the positions of items. For instance, the atoms \( \text{pat}(1,a), \text{pat}(2,b), \) and \( \text{pat}(2,c) \) stand for the (frequent) pattern \( \langle a \langle bc \rangle \rangle \) of the database given in Listing 1.

A complete specification of atoms used in our ASP encodings along with their meanings is given in Table 1. In more detail, the rule in Line 1 of Listing 2 provides items occurring in a given database D. Lines 3 and 4 provides the slots for the itemset of a pattern with a varying length \( m \in [\text{max}] \). Line 7 evaluates the concrete length of the candidate pattern. The (choice) rule in line 5 allows for picking a subset of \( \Sigma \) per slot to build a pattern. As a result, the rules from lines 1 to 5 establish a unique representation for a candidate pattern whose frequency remains to be checked.

To this end, the rules from Line 9 to 12 traverse each sequence \( t = (t_1 \ldots t_n) \) in D to derive atoms of the form \( \text{emb}(t,p,x) \) which represent all the embeddings of the pattern in a sequence. As noted in Table 1 for \( p \in [n] \) and \( x \in [\text{max}] \), such an atom expresses that it exists \( (t_1 \ldots t_i) \subseteq t \) such that \( p_i \leq t_i \) for all \( i \in [p-1] \) and \( p_o \leq t_p \) where \( p \in [n] \) and \( x \in [\text{max}] \). Thus, it also means that \( \langle p_1 \ldots p_o \rangle \subseteq \langle t_1 \ldots t_p \rangle \). Line 9 locates the occurrences of the first itemset in the sequence. The expression \( \text{seq}(t,p,e) \) : \( \text{pat}(1,e) \) ensures that every \( e \in p_1 \) must hold at position \( p \) of the sequence \( t \). Line 10 finds out locations for \( p_o \), the \( x \)-th itemset of the pattern, from a position \( q \) such that \( q \geq e_i \) for all \( i \in [p-1] \), where \( (t_1 \ldots t_{p_o}) \) is an embedding of \( \langle p_1 \ldots p_{o-1} \rangle \). The atoms \( \text{emb}(t,p,x) \) are further inspected, line 14, to determine the cover of a candidate pattern \( p \). In these lines, the sequence is covered when a complete embedding has been built, i.e. up to the length of the pattern. The frequency of \( p \) is then checked in Line 16, where the cardinality constraint \( k \{ \text{cover}(t) \} \) over atoms \( \text{cover}(t) \), signaling \( a \subseteq t \) expresses that at least the frequency threshold \( k \) many sequences \( t \) must include \( a \).

Example 1. For the database in Listing 2, the pattern \( \langle a \langle bc \rangle \rangle \) leads to atoms \( \text{emb}(3,1,1) \), \( \text{emb}(3,1,3) \) and \( \text{emb}(3,2,2) \) among which \( \text{emb}(3,2,2) \) in turn yields \( \text{cover}(3) \) for the sequence \( \langle a \langle bc \rangle \rangle \) denoted by 3. There is not any sequence for which a complete embedding can be built, and \( \langle a \langle bc \rangle \rangle \) turns out to not be frequent, given the threshold \( k = 2 \).

4 Negative sequential patterns

In this section, we introduce the notion of negative sequential pattern by giving their syntax and semantics borrowed from [3].
Table 1. Atoms representing a pattern $p = (p_1)_{1 \leq i \leq m}$, where $1 \leq m \leq \text{max}$, and a given multiset $D$ of sequences $t = (t_j)_{1 \leq j \leq n}$

<table>
<thead>
<tr>
<th>Atom</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>slot(x)</td>
<td>$1 \leq x \leq m$ refers to the position $x$ of an item $s_i$ in $p$</td>
</tr>
<tr>
<td>pat(x,e)</td>
<td>$e$ is the item at position $x$ in $p$, where $x \in [m]$</td>
</tr>
<tr>
<td>item(e)</td>
<td>$e$ belongs to some $t \in D$</td>
</tr>
<tr>
<td>cover(t)</td>
<td>$(p_1 \ldots p_m) \subseteq (t_1 \ldots t_n)$, that is, $p \subseteq t$</td>
</tr>
<tr>
<td>emb(t,p,x)</td>
<td>$\exists (e_i)<em>{[p-1]}$ such that $p_i \subseteq t</em>{e_i}$ for all $i \in [p-1]$ and $p_n \subseteq t_p$, where $p \in [n]$ and $x \in [m]$.</td>
</tr>
</tbody>
</table>

Definition 1 (Negative sequential patterns (NSP)). A negative sequential pattern $p = (p_1 \ldots q_1 \ldots q_n \ldots p_m)$ is a finite sequence where $p_i \in 2^S \setminus \emptyset$ for all $i \in [n]$ and $q_i \in 2^S$ for all $i \in [n-1]$. $p^+ = (p_1 \ldots p_n)$ is called the positive part of the NSP.

It can be noticed that Definition 1 introduces syntactic limitations on negative sequential patterns that are commonly encountered in the state of the art [13]: 1) a pattern can neither start or finish by a negative itemset, 2) a pattern cannot have two successive negative itemsets. But a pattern can have successive positive itemsets considering that a negative itemset can be empty.

Example 2. This example illustrates the notations introduced in Definition [7]. Consider $\Sigma = \{a,b,c,d\}$ and $p = \langle a \sim (bc) (ad) \rangle$ and $d \sim (ab) d \rangle$. Let $p_1 = \{a\}$, $p_2 = \{ad\}$, $p_3 = \{d\}$ and $q_1 = \{bc\}$, $q_2 = \emptyset$, $q_3 = \{ab\}$. $p^+ = \langle a (ad) d \rangle$.

4.1 Semantics of negative sequential patterns

The semantics of negative sequential patterns relies upon negative containment: a sequence contains pattern $p$ if $s$ contains a subsequence $s'$ such that every positive itemset of $p$ is included in some itemset of $s'$ in the same order and for any negative itemset $\sim q_i$ of $p$, $q_i$ is not included in any itemset occurring in the sub-sequence of $s'$ located between the occurrence of the positive itemset preceding $q_i$ in $p$ and the occurrence of the positive itemset following $q_i$ in $p$.

It turns out that the notion of negative containment has different semantics depending on the definition of “not included” itemsets and on the itemsets of the sequence to consider.

We introduce two relations comparing two itemsets $P \in 2^S \setminus \emptyset$ and $I \in 2^S$:

- partial non inclusion: $P \not\subseteq G I \leftrightarrow \exists e \in P, e \notin I$
- total non inclusion: $P \not\subseteq D I \leftrightarrow \forall e \in P, e \notin I$

Partial non-inclusion means that $P \subseteq I$ is non-empty while total non-inclusion means that $P$ and $I$ are disjoint. By convention, $\emptyset \not\subseteq D I$ and $\emptyset \not\subseteq G I$ for all $I \subseteq \Sigma$.

In the sequel, we denote the general form of itemset non-inclusion by the symbol $\not\subseteq$, meaning either $\not\subseteq G$ or $\not\subseteq D$.

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:

$$D = \{ s_1 = ((bc) f a) \quad s_2 = ((bc) (ef) a) \quad s_3 = ((bc) (df) a) \quad s_4 = ((bc) (ef) a) \quad s_5 = ((bc) (cdef) a) \}$$

Table 2 compares the support of patterns under the two semantics of itemset non-inclusion. Since the positive part of $p_2$ is in $s_2$, $p_2$ occurs in the sequence iff $(cd) \not\subseteq G (ce)$. As for total non-inclusion, it is false that $(cd) \not\subseteq D (ce)$ because $c$ occurs in $(ce)$, and thus $p_2$ does not occur in $s_2$. As for partial non-inclusion, it is true that $(cd) \not\subseteq G (ce)$, because $d$ does not occur in $(ce)$, and thus $p_2$ occurs in $s_2$.

Table 2. Lists of sequences in $D$ supported by negative patterns $(p_i)_{i=1..4}$ under the total and partial non-inclusion relations. Each pattern has the form $(b \sim q_i)$, where $q_i$ are itemsets such that $q_i \subseteq q_{i+1}$.

<table>
<thead>
<tr>
<th>Atom</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>slot(x)</td>
<td>$1 \leq x \leq m$ refers to the position $x$ of an item $s_i$ in $p$</td>
</tr>
<tr>
<td>pat(x,e)</td>
<td>$e$ is the item at position $x$ in $p$, where $x \in [m]$</td>
</tr>
<tr>
<td>item(e)</td>
<td>$e$ belongs to some $t \in D$</td>
</tr>
<tr>
<td>cover(t)</td>
<td>$(p_1 \ldots p_m) \subseteq (t_1 \ldots t_n)$, that is, $p \subseteq t$</td>
</tr>
<tr>
<td>emb(t,p,x)</td>
<td>$\exists (e_i)<em>{[p-1]}$ such that $p_i \subseteq t</em>{e_i}$ for all $i \in [p-1]$ and $p_n \subseteq t_p$, where $p \in [n]$ and $x \in [m]$.</td>
</tr>
</tbody>
</table>

Now, we formulate the notions of sub-sequence, non-inclusion and absence by extending the concept of embedding (see Section 2.1) to negative patterns.

Definition 2 (Strict and soft embeddings of negative patterns). Let $s = (s_1 \ldots s_n)$ be a sequence and $p = (p_1 \sim q_1 \ldots \sim q_{m-1} \sim p_m)$ be a negative sequential pattern.

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

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

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

- $p_i \subseteq s_{e_i}$ for all $i \in [m]$
- $q_i \not\subseteq s_j$, for all $i \in [e_i + 1, e_{i+1} - 1]$ while a strict embedding consider them as a whole.

Example 3 (Itemset absence semantics). Let $p = (a \sim (bc) d)$ be a pattern and consider four sequences as follows:

- By an increasing tuple $e$, we mean a tuple such that $e_i < e_{i+1}$ (in particular, repetitions are not allowed).
Lemma 1. In the case that $\mathcal{G}_s \subseteq \mathcal{G}_D$, e is a soft-embedding iff e is a strict-embedding.

Lemma 2. Let $p = (p_1 \neg q_1 \ldots \neg q_{n-1} p_n) \in \mathcal{N}$ such that $|q_i| \leq 1$ for all $i \in [n - 1]$, then e is a soft-embedding iff e is a strict-embedding.

At this point, we have exhibited four semantics for negative pattern embeddings. Lemmas 1 and 2 concluded to the equivalence of soft and strict-embeddings when the itemset non-inclusion relation is $\mathcal{G}_D$.

In [8], we proposed notations to disambiguate syntactically the different types of negation ($\neg(a_1, \ldots, a_n)$ or $\neg\{a_1, \ldots, a_n\}$). Indeed, items $b$ and $c$ occur between occurrences of $a$ and $d$ in $s_1$ and $s_2$. Considering total non-inclusion ($\mathcal{G}_s := \mathcal{G}_D$) and either type of embeddings, the absence of an itemset is satisfied if any of its items is absent. Hence, $p$ occurs only in $s_4$.

Let us now give some lemmata about the relations between the different types of embedding:

Lemmas 1 and 2 concluded to the equivalence of soft and strict-embeddings when the itemset non-inclusion relation is $\mathcal{G}_D$.

In [8], we proposed notations to disambiguate syntactically the different types of negation ($\neg(a_1, \ldots, a_n)$ or $\neg\{a_1, \ldots, a_n\}$) or $\neg\{a_1, \ldots, a_n\}$. $\mathcal{N}$ denotes the interval of integers $[l + 1, u - 1]$.

- $\neg(a_1, \ldots, a_i)$ is evaluated as $\{a_1, \ldots, a_i\} \subseteq \mathcal{G}_s s_j, \forall j \in [l_i, e_i+1]$ for all $i \in [m - 1]$ (2).

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 $a_1, \ldots, a_i$.

- $\neg\{a_1, \ldots, a_i\}$ is evaluated as $\{a_1, \ldots, a_i\} \subseteq \mathcal{G}_s \bigcup_{j \in [l_i, e_i+1]} s_j$ for all $i \in [m - 1]$ (3).

Intuitively, we check that there exists some item in $a_1, \ldots, a_i$ 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 $p_{i+1}$).

- $\neg a_1, \ldots, a_i$ is evaluated as $\{a_1, \ldots, a_i\} \subseteq \mathcal{G}_s s_j, \forall j \in [l_i, e_i+1]$ for all $i \in [m - 1]$ (4).

Intuitively, we check that every item in $a_1, \ldots, a_i$ 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 $p_{i+1}$).

4.2 Multiple occurrences

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?

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

The strong containment relation states that a negative pattern $p$ occurs in a sequence $s$ if there exists at least one occurrence of the positive part of pattern $p$ in sequence $s$ and every such occurrence satisfies the negative constraints; the strong containment relation states $p$ occurs in a sequence $s$ if there exists at least one occurrence of the positive part of pattern $p$ in sequence $s$ and at least one of these occurrences satisfies the negative constraints.

Example 4 (Strong vs weak occurrence semantics). Let $p = \langle a b \neg c d \rangle$ be a pattern, $s_1 = \langle a b c e d \rangle$ and $s_2 = \langle a b c a d e b d \rangle$ be two sequences. Thus, $p^+ = \langle a, d \rangle$ occurs once in $s_1$, hence there is no difference for occurrences of $p$ in $s_1$ under the two semantics. However, $p^+$ occurs four times in $s_2$ 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 (¬) but the last two occurrences do. Under the weak occurrence semantics, pattern $p$ occurs in sequence $s_2$ whereas it fails to do so under the strong occurrence semantics.

5 ASP encodings of negative sequential patterns

In the following, we introduce new facts to encode negations in negative sequential patterns. The $\text{pat}(x, e)$ atom now represents the positive part of a negative sequential pattern. The negated itemsets are encoded by terms of three predicates: $\text{negpat}(2, \neg \cdot)$ and $\text{csnegpat}(2, \neg \cdot)$. For instance, the pattern $\langle a \neg|b|a\neg|bd|c \rangle$ is encoded by the facts $\text{pat}(1, a)$, $\text{pat}(2, a)$, $\text{pat}(3, c)$, $\text{negpat}(1, b)$, $\text{csnegpat}(2, b)$ and $\text{csnegpat}(2, d)$.

The following choice rule illustrates the generation of negative itemsets of kind $\neg|\cdot|$. Similar rules yield the negative part of a candidate pattern for the two other kinds of negations.

\[
\text{negpat}(X, E, \text{item}(E), \text{not pat}(X, E)) \leftarrow \\
\text{slot}(X), X < L, \text{patlen}(L).
\]

Note that we assume that it is not possible to have several different kinds of negated itemsets at the same position. For instance, pattern $\langle a \neg|b|\neg|c|d \rangle$ is not possible.

In the next section, we give the rules for negative constraints. A negative constraint denotes a constraint upon an embedding $(e_i)_{i \in [n]}$ of the positive part that is expressed by a negated itemset in the candidate negative pattern. If all negative constraints are satisfied, then $(e_i)_{i \in [n]}$ is an embedding of the negative pattern. Then, it is further used to evaluate the containment of the candidate pattern $p$ in a sequence $t$.

The core of the encoding relies on the construction of an embedding. We adapt the definition of the predicate $\text{emb}/3$ to negative sequential patterns as follow: let $p = \langle p_1, \neg q_1 \ldots p_m \rangle$ be a negative sequential pattern, $t = \langle t_1 \ldots t_n \rangle$ be a sequence, and $\nu$ a Boolean value, then $\text{emb}(t, p, \nu)$ means that $\exists (e_i)_{i \in [x - 1]}$ such that $x \in [m], \rho \in [n]$ and

1. $p_i \subseteq t_{e_i}$ for all $i \in [x - 1]$,
2. $p_i \subseteq t_{e_{i+1}}$,
3. $\exists y \in [x - 1]$ s.t. $q_y$ is not satisfied for the partial embedding $(e_i)_{i \in [x - 1]}$ if $\nu$ is true,
4. $\forall y \in [x - 1]$ s.t. $q_y$ is satisfied for the partial embedding $(e_i)_{i \in [x - 1]}$ if $\nu$ is false.

Note that “$q_y$ is not satisfied for the partial embedding $(e_i)_{i \in [x - 1]}$” refers to the semantics of negation (see previous section) whatever the kind of occurrence (weak or strong).
\begin{verbatim}
1 emb(T,P,1,0) :- seq(T,P,E):pat(1,E); seq(T,P,._).
3 emb(T,P,X+1,1) :- emb(T,Q,X,1); Q<P;
   seq(T,P,E):pat(X+1,E); seq(T,P,._);
5 patlen(L); X<L.
7 emb(T,P,X+1,0) :- emb(T,Q,X,0); seq(T,P,E):pat(X+1,E); seq(T,P,._); P>Q;
   #count{ R : not seq(T,R,Ep), csnegpat(X,Ep), R=\langle Q+1..P-1\rangle }=(P-\langle Q\rangle-1);
9 patlen(L); X<L.
11 emb(T,P,X+1,1) :- emb(T,Q,X,0);
13 seq(T,P,E):pat(X+1,E); seq(T,P,._); P>Q;
14 #count{ R : not seq(T,R,Ep), csnegpat(X,Ep), R=\langle Q+1..(P-1)\rangle }=(P-\langle Q\rangle-1);
15 patlen(L); X<L.

Listing 3. Encoding for embedding with \texttt{-}(-)\texttt{csnegpat}/2 negative itemsets

\begin{verbatim}
1 emb(T,P,X+1,1) :- emb(T,Q,X,0); Q<P;
2 seq(T,P,E):pat(X+1,E); seq(T,P,._);
3 #count{ R : seq(T,R,Ep), negpat(X,Ep), Q<R, R<P }=0;
4 patlen(L); X<L.
6 emb(T,P,X+1,0) :- emb(T,Q,X,0); Q<P;
7 seq(T,P,E):pat(X+1,E); seq(T,P,._);
8 #count{ R : seq(T,R,Ep), negpat(X,Ep), Q<R, R<P }=0;
9 patlen(L); X<L.

Listing 4. Encoding for embedding with \texttt{-}(-)\texttt{negpat}/2 negative itemsets
\end{verbatim}

\nu\) is a Boolean value that stores the information for an embedding of the positive part that these embedding satisfies or not the negative constraints.

With this definition it happens that if there is an atom of the embedding \texttt{emb}(t,p,l,\nu) where \(l\) is the size of the positive part of the pattern whatever the \(\nu\) value, then the positive part occurs in the sequence \(t\). If \(\nu\) is false, this means that there exists an embedding for the negative sequential pattern, thus \(p\) weakly occurs in \(t\). If there is no embedding such that \(\nu\) is true, then it means that no embedding of the positive part is an embedding of the negative sequential pattern. Thus \(p\) strongly occurs in \(t\) if it is not the case. The following rules encode the rules of pattern covering for the two different definitions of weak and strong occurrences of a candidate pattern.

\texttt{weakcover}(T) :- patlen(L); emb(T,\_,L,0).
\texttt{strongcover}(T) :- patlen(L); emb(T,\_,L,\_); #count( T : emb(T,\_,L,\_1) )=0.

Then, depending on the kind of covering considered, line 16 of Listing 2 is adapted to the minimal occurrence number constraint.

The remainder of this section details the encoding for each semantics of negative itemsets.

5.1 Encoding of \texttt{-}(-) negative constraints

Listing 3 gives the encoding yielding the \texttt{emb}/4 atoms considering patterns with \texttt{-}(-)\texttt{csnegpat}/2 negative itemsets.

In this encoding, the Line 1 states that \(\epsilon_1\) of an embedding is a position \(p\) in the sequence \(t\) such \(p_1 \subseteq t_p\). For the first pattern itemset, there are no negative constraints at previous position. Thus, \(\nu\) is false (0).

The rule in lines 3-5 is similar to the rule in lines 10-12 of Listing 2. They yield the \(\epsilon_{p+1}\) of an embedding of the positive part of the pattern where \((\epsilon_i)_{i \in |x|}\) is a embedding of \((p_1 \sim q_1 \ldots p_e)\) s.t. there exists \(y \in [x-1]\) s.t. \(q_y\) is not satisfied (\(\nu\) is true). In this case, \((\epsilon_i)_{i \in |x|}\) is an embedding of \((p_1 \sim q_1 \ldots p_e)\) that neither satisfies all the negative constraints. Thus, the truth value of \(\nu\) is propagated.

The rules in lines 7-15 consider the case of an embedding \((\epsilon_i)_{i \in |x|}\) of the negative pattern \((p_1 \sim q_1 \ldots p_e)\) such that all the negative constraints are satisfied. Then, there are two cases for the embedding \((\epsilon_i)_{i \in |x|}\). If there exists a position \(p\) in the sequence \(t\) such that \(p_{e+1} \subseteq t_p\) and the negative constraint \(q_e\) is satisfied, then the atom \texttt{emb}(t,p,l,\nu) is yielded with \(\nu\) set to false (0). That is, the rule in lines 7-10. But, if there exists a position \(p\), \(p > \epsilon_1\) in the sequence \(t\) such that \(p_{e+1} \subseteq t_p\) and the negative constraint \(q_e\) is not satisfied, then the atom \texttt{emb}(t,p,l,\nu) is yielded with \(\nu\) set to true (1). That is, the rule in lines 12-15. Note that these two rules could be factorised to avoid redundancies. For sake of readability, we prefer to present this redundant syntax.

Let us now detail how to evaluate the negative constraint \(q_e\) (i.e. the set of atoms \texttt{csnegpat}(x,e)). According to Equation 2, the negative constraint \(q_e\) is satisfied iff \(\forall r \in [\epsilon_1,\epsilon_{e+1}]\ [e \notin q_e, e \notin t_r\). It is equivalent to evaluate whether \([\{r \in [\epsilon_1,\epsilon_{e+1}]\mid e \notin q_e, e \notin t_r\} \subseteq \epsilon_1 - \epsilon_1 + 1]\). This latter constraint is encoded line 9 and its opposite is encoded line 14. The \#count is an aggregate to compute the number of different \(R\) values satisfying conditions after the colon.

5.2 Encoding of \texttt{-}(-) negative constraints

The negative constraints \(q_e\) with \texttt{-}(-)\texttt{negpat}/2 is satisfied iff \(\forall r \in [\epsilon_1,\epsilon_{e+1}]\ [e \notin q_e, e \notin t_r\) (see Equation 4). This means that it is not satisfied iff \(\{r \in [\epsilon_1,\epsilon_{e+1}]\mid e \notin q_e, e \notin t_r\}\) is empty. Listing 4 yields atoms of predicate \texttt{emb}/4 with the same principle of Listing 5. The lines 3 or 8 evaluate \([\{r \in [\epsilon_1,\epsilon_{e+1}]\mid e \notin q_e, e \notin t_r\}\)
5.3 Encoding of ¬{·} negative constraints

Finally, the negative constraint $q_v$ with ¬{·}/cnegpat/2 is satisfied iff $\exists e \in q_v, e \not\in \bigcup_{t \in \tau} (e, t)$. See Equation 3. This means that it is not satisfied if $\forall e \in q_v, e \in \bigcup_{t \in \tau} (e, t)$, or in other words that $|\{e \in q_v | e \in \bigcup_{t \in \tau} (e, t)\}| = |q_v|$. Listing 5 yields atoms of predicate emb/4 with the same principle of Listing 3. Line 1 evaluates the size of negative itemsets. negislen(\(x, n\)) states that $|q_v| = n$. Lines 6 and 12 evaluate $\{e \in q_v | e \in \bigcup_{t \in \tau} (e, t)\}$. It is compared to $|q_v|$ to assign the truth value to $\nu$ for $\epsilon_{\epsilon+1}$ (if exists, i.e., $p_{x+1} \subseteq t_{x+1}$).

6 Experiments and results

The previous section exhibited the ease of modeling of the different semantics of negative sequential patterns in order to extract the frequent ones. It is worth noticing that only few of these semantics have properties to design algorithms that are both efficient and complete \[8\]. More precisely, the anti-monotonicity property holds with weak occurrences of $\neg | \cdot |$, and weaker versions of anti-monotonicity hold with the strong occurrences of $\neg | \cdot |$.

The question investigated in these experiments is whether the solver takes advantage of these properties to achieve the mining task within a lower computation time.

6.1 Experimental setting

We conducted experiments on synthetic sets of sequences generated with the generator of \[2\], without hidden patterns (pure random sequences).

The usage of synthetic sets of sequences allows us to control and analyze the most important characteristics of data wrt the computing time. To be more precise, we randomly generated sets of 100 sequences. Each sequence contains 10 itemsets. Each itemset is of size 2. Items are randomly generated according to a Gaussian law (some items are more frequent than others) over a vocabulary of 20 items. For each data point, we then group average computing time over five similar sets of sequences.

All the experiments have been using the clingo solver\[^5\] on a desktop computer with enough memory to prevent from using cache memory and ran with a single thread. A timeout was set up to 5 minutes.

For sake of better interpretability of the results and of computation time, the negative sequential patterns contain only one kind of negated itemset ($\neg | \cdot |, \neg \cdot$ or $\neg \cdot$), the length of the positive part of a pattern is at most 3, and the size of negated itemsets is limited to 2 items.

Note that we do not compare our approach with dedicated algorithms, e.g., \[10\][17], that are undoubtedly significantly more efficient for this task. We remind that declarative pattern mining aims at achieving reasonable computation time for mid-sized set of examples. Thus, it can be used to inject some expert knowledge in the mining process to reduce the number of patterns.

6.2 Results

Figure 1 compares the number of patterns that are extracted through the different semantics of negative patterns. Note that it is very interesting to have this study as no algorithm can extract a complete set of frequent negative sequential patterns for all negation semantics.

The figures take into account only the number of patterns for runs ended before timeout (5 minutes). It explains that there are only few results for $\neg \cdot$ which is often out of the timeout.

The results illustrated in Figure 1 are consistent with theoretical results \[8\]. First, we notice that there are more patterns extracted with weak-occurrences than with strong-occurrences. For identical settings, there are in average 2.80 ± 2.13 times more patterns with weak-occurrences than with strong-occurrences. Second, $\neg | \cdot |$ is the strongest negative constraint after $\neg \cdot$ and finally $\neg \cdot$ is the weakest. As a consequence, there are more extracted patterns for $\neg \cdot$, than for $\neg | \cdot |$. The number of patterns for $\neg \cdot$ is in-between when available.

Figure 2 compares computing times of ASP-based negative sequential pattern mining with different semantics of negative patterns.

We can first notice that extracting negative sequential patterns with $\neg \cdot$ lacks efficiency and the mining tasks does not finish before timeout in most cases.

Then, despite the larger number of occurrences of weak-occurrences, evaluating weak occurrences is more efficient that evaluating the strong ones. It is unsurprising. Indeed, as soon as a weak-occurrence has been found, the atom $cover(t)$ is proved to be true without necessarily exploring the other embeddings of the positive part of the pattern. For strong occurrences, it is mandatory to evaluate negative patterns for every embedding of the positive part of the candidate pattern.

\[^5\] clingo version 5.3.1
The surprising result of this experiment is in the comparison between $\neg(\cdot)$ and $\neg|\cdot|$. It turns out that mining negative sequential patterns with $\neg(\cdot)$ is significantly more efficient than with $\neg|\cdot|$. This result is enhanced by the fact that there are many more patterns to extract with $\neg(\cdot)$ than with $\neg|\cdot|$. So there are more patterns but more efficiently extracted. The opposite outcome was expected due to our theoretical results that exhibited anti-monotonicity for $\neg|\cdot|$, but not for $\neg(\cdot)$.

7 Conclusion

This article extends the framework of mining sequential patterns with Answer Set Programming to negative sequential patterns. We encoded the three different semantics of absent events proposed in [8]. The first interest of the approach is to evaluate the impact of the choice of one of the semantics on the set of extracted patterns. In this article, we focused on the size of this set. Deeper experiments on real datasets would be interesting to identify the differences between these sets of patterns.

Our experiments on synthetic data show that the kind of negated itemset for which the encoding is the most efficient is not the expected one. Indeed, we expected to have better efficiency with $\neg|\cdot|$. A possible explanation of this observation may be that there are also anti-monotonicity properties of $\neg(\cdot)$ for partial orders that are not usually investigated for pattern mining algorithms [10]. More investigation is required to confirm this intuition. But it encourages us to think that there are potentially better encodings for $\neg|\cdot|$. One of our perspectives is to investigate saturation programming techniques that may be suitable and efficient to encode weak-occurrences of negative sequential patterns.

In addition, these encodings may be used to characterize what patterns can still be mined by incomplete mining heuristics. Wang et al. [17] reviewed such algorithms but they do not characterise the extracted patterns among the complete set of patterns. Comparing their outputs with the outputs of our encodings makes it possible to investigate this question.

REFERENCES


