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

Thomas Guyet

Inria – AIstroSight

CNIA 2023

Outline

Sequential pattern mining

- A **sequence** $\boldsymbol{s}$ is an ordered set of events (or **itemsets**)

s_1	$\langle a(abc)(ac)d(cf)\rangle$
s_2	$\langle (ab)c(bc)(ae)(ad)\rangle$
s_3	$\langle eg(af)cbc(de)\rangle$

Sequential pattern mining

- A **sequence** $\boldsymbol{s}$ is an ordered set of events (or **itemsets**)
- A **sequential pattern** is a subsequence
  - containment relation: $\boldsymbol{p} \preceq \boldsymbol{s}$
    - → inclusion of itemsets
    - → gaps are allowed
  - example:
    - pattern $\boldsymbol{p} = \langle a(bc)d \rangle$
    - embedding: mapping of a pattern on a sequence ($\langle 1,2,4 \rangle$, $\langle 1,3,5 \rangle$)

$\boldsymbol{s}_1$	$\langle a(abc)(ac)d(cf)\rangle$
$\boldsymbol{s}_2$	$\langle (ab)c(bc)(ae)(ad)\rangle$
$\boldsymbol{s}_3$	$\langle eg(af)cbc(de)\rangle$

Sequential pattern mining

- Let $\mathcal{D}$ be a set of sequences,
  - **Frequent pattern mining**: Given a support threshold σ, find the complete set of sequential patterns with support above σ
  - the **support** of pattern $\boldsymbol{p}$ in $\mathcal{D}$ is the number of sequences in $\mathcal{D}$ that contain $\boldsymbol{p}$:

$$supp(\boldsymbol{p}) = |\{\boldsymbol{s} \in \mathcal{D} | \boldsymbol{p} \preceq \boldsymbol{s}\}|$$

s_1	$\langle a(abc)(ac)d(cf)\rangle$
s_2	$\langle (ab)c(bc)(ae)(ad)\rangle$
s_3	$\langle eg(af)cbc(de)\rangle$

$$supp(\langle a(bc)d\rangle) = 2$$

Motivation for Negative Sequential Patterns

Positivism of frequent sequential pattern mining

Frequent pattern mining algorithms extract only patterns as subsequences that actually occur!

Problem with frequent sequential pattern mining

- Dataset with hidden frequent patterns
  - $\text{sympt}_1 \rightsquigarrow \ldots \rightsquigarrow \text{sympt}_n \rightsquigarrow \text{disease}$
  - $\text{sympt}_1 \rightsquigarrow \ldots \rightsquigarrow \text{drug} \rightsquigarrow \ldots \rightsquigarrow \text{sympt}_n$

  → disease appears only when no drug has been taken
- extracted pattern

  → $\text{sympt}_1 \rightsquigarrow \ldots \rightsquigarrow \text{sympt}_n$

  → **not really useful for our problem**

What kind of pattern would be interesting?

→ patterns that may highlight the **absence** of an item (the so-called *negative items*)

- $\text{sympt}_1 \rightsquigarrow \ldots \rightsquigarrow$ **no** drug $\rightsquigarrow \ldots \rightsquigarrow \text{sympt}_n \rightsquigarrow \text{disease}$

Negative sequential patterns in the State of the Art

- Few algorithms extract negative sequential patterns
  - eNSP [CDZ16] and its variants
  - NegGSP [ZZZC09]
  - Gong et al [GLD15]
  - PNSP [HLC08]
  - NegPSpan [GQ20]
- Analysis of the state of the art [BG20]
  - **State of the art algorithms do not extract the same patterns**
  - There are **several semantics for patterns with negation** in sequences of itemsets

Our research questions

1. Are there “intuitive” semantics for patterns with negation?
2. Do the “intuitive” semantics correspond to those actually used by one of the algorithms?
3. What recommendations about the use of patterns with negations?

Methodology: survey about the perception of negation in NSP

1. Identification of alternative interpretations of the negation
   - We adhere to the analysis of Besnard and Guyet [BG20]
   - ⇒ $2^3 = 8$ possible semantics: two alternative perceptions for 3 dimensions
2. Design of the survey
   - Should be answered by people without preliminary knowledge about pattern mining
   - Characterization of interviewed
   - Attempt to capture additional bias
   - Anonymity
3. Collection of answers
   - Broadcast on national and international mailing lists (in DM and AI)
   - Broadcast to people (personal circles) without preliminary knowledge in data science
   - → Attempt to have a broad range of people (*not assessed*)
4. Analysis of the survey answers

Outline

Negative patterns: a syntactic definition [BG20]

We take $\mathcal{I}$ to denote the set of possible items.

Definition (Negative sequential patterns (NSP))

A negative pattern $\boldsymbol{p} = \langle p_1 \ \neg q_1 \ p_2 \ \neg q_2 \ \dots \ p_{n-1} \ \neg q_{n-1} \ p_n \rangle$ 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 $i \in [n-1]$.

Syntactic limitations on negative sequential patterns

- an NSP can neither start nor finish with a negative pattern,
- an NSP cannot have two successive negative itemsets,
- an NSP cannot specify positive and negative items in the same position.

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

Semantics of NSP

!!! Spoiler Alert !!!

Do not listen the end of this talk if you want contribute to the survey !

`https://tinyurl.com/NegativePatternsSurvey`

Semantics of NSP [BG20]

- The **containment relation** between an NSP $\boldsymbol{p}$ and a sequence $\boldsymbol{s}$ defines the **semantics** of NSPs
  ↪ Different containment relations lead to different support measures for a pattern, and thus negative sequential pattern mining algorithms does not extract the same pattern set.

8 possible semantics depending on how to consider

- partial vs total itemset non-inclusion
- soft vs strict embeddings
- weak vs strong occurrences

Partial/total itemset non-inclusion

$$\boldsymbol{p}_2 = \langle b \neg (cd) a \rangle$$

$\mathcal{D}_1 =$

$\boldsymbol{s}_1 = \langle b\ f\ a \rangle$
$\boldsymbol{s}_2 = \langle b\ (cf)\ a \rangle$
$\boldsymbol{s}_3 = \langle b\ (df)\ a \rangle$
$\boldsymbol{s}_4 = \langle b\ (ef)\ a \rangle$
$\boldsymbol{s}_5 = \langle b\ (cdef)\ a \rangle$

Partial non-inclusion ($\not\subseteq_G$)
$\boldsymbol{s}_1 = \langle b\ f\ a \rangle$
$\boldsymbol{s}_2 = \langle b\ (cf)\ a \rangle$
$\boldsymbol{s}_3 = \langle b\ (df)\ a \rangle$
$\boldsymbol{s}_4 = \langle b\ (ef)\ a \rangle$
~~$\boldsymbol{s}_5 = \langle b\ (cdef)\ a \rangle$~~

Total non-inclusion ($\not\subseteq_D$)
$\boldsymbol{s}_1 = \langle b\ f\ a \rangle$
~~$\boldsymbol{s}_2 = \langle b\ (cf)\ a \rangle$~~
~~$\boldsymbol{s}_3 = \langle b\ (df)\ a \rangle$~~
$\boldsymbol{s}_4 = \langle b\ (ef)\ a \rangle$
~~$\boldsymbol{s}_5 = \langle b\ (cdef)\ a \rangle$~~

Soft/strict-embeddings

$\boldsymbol{p}_3 = \langle a \neg (bc) d \rangle$

$\mathcal{D}_2 =$

$\boldsymbol{s}_1 = \langle a\ c\ b\ e\ d \rangle$
$\boldsymbol{s}_2 = \langle a\ (bc)\ e\ d \rangle$
$\boldsymbol{s}_3 = \langle a\ b\ e\ d \rangle$
$\boldsymbol{s}_4 = \langle a\ e\ d \rangle$

soft-embedding $\circ$
$(\forall j \in [e_{i-1}+1, e_{i+1}-1],\ q_i \nsubseteq s_j)$:

$\boldsymbol{s}_1 = \langle a\ c\ b\ e\ d \rangle$
$\sout{\boldsymbol{s}_2 = \langle a\ (bc)\ e\ d \rangle}$
$\boldsymbol{s}_3 = \langle a\ b\ e\ d \rangle$
$\boldsymbol{s}_4 = \langle a\ e\ d \rangle$

strict-embedding, $\bullet$
$(q_i \nsubseteq \bigcup_{j \in [e_{i-1}+1, e_{i+1}-1]} s_j)$:

$\sout{\boldsymbol{s}_1 = \langle a\ c\ b\ e\ d \rangle}$
$\sout{\boldsymbol{s}_2 = \langle a\ (bc)\ e\ d \rangle}$
$\boldsymbol{s}_3 = \langle a\ b\ e\ d \rangle$
$\boldsymbol{s}_4 = \langle a\ e\ d \rangle$

Weak/strong occurrences

$\boldsymbol{p}_4 = \langle ab\neg cd \rangle$

$\mathcal{D}_3 =$

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

weakly-occur, $\preceq$ (*there exists*):

$s_1 = \langle a\ b\ e\ d \rangle$
$s_2 =$ ~~$\langle a\ b\ c\ d\ e\ b\ d \rangle$~~, $\langle a\ b\ c\ d\ e\ b\ d \rangle$
$s_3 = \langle a\ e\ d\ b\ e\ d\ d \rangle, \langle a\ e\ d\ b\ e\ d\ d \rangle$
~~$s_4 = \langle a\ e\ d\ b\ c\ e\ d \rangle$~~

strongly-occur, $\sqsubseteq$ (*for each positive*):

$s_1 = \langle a\ b\ e\ d \rangle$
~~$s_2 = \langle a\ b\ c\ d\ e\ b\ d \rangle$~~, $\langle a\ b\ c\ d\ e\ b\ d \rangle$
$s_3 = \langle a\ e\ d\ b\ e\ d\ d \rangle, \langle a\ e\ d\ b\ e\ d\ d \rangle$
~~$s_4 = \langle a\ e\ d\ b\ c\ e\ d \rangle$~~

Outline

Overall organisation of the survey

1. Evaluation of the level of knowledge in the domain of pattern mining and/or logic
   - → self-assessment of the background knowledge about pattern mining
   - → identification of specific skills (computer science, data science, logic)
2. Preliminary check of the understanding of the basics of (positive) sequential patterns
   - → One verification question: **the user can not access the next questions until s/he correctly answered it**
3. 5 questions about the semantics
   - → scope of the negation
   - → three dimensions of NSP's semantics: non-inclusion, embeddings, occurrences
   - → *(one question about the strength of negation vs multiplicity)*

- More details about the questions are provided in the article
- Survey: `https://tinyurl.com/NegativePatternsSurvey`

Example of question: multiple occurrences

According to you, what are the sequences that contain the pattern $\boldsymbol{p} = \langle b \; \neg e \; f \rangle$?

id	*Sequence*
$\boldsymbol{o}_0$	$\langle b \; a \; f \; d \; b \; d \; f \rangle$
$\boldsymbol{o}_1$	$\langle b \; a \; f \; d \; e \; b \; d \; f \rangle$
$\boldsymbol{o}_2$	$\langle d \; b \; e \; c \; a \; d \; f \; b \; d \; e \; f \rangle$
$\boldsymbol{o}_3$	$\langle b \; a \; f \; b \; a \; e \; f \rangle$

- The user is invited to decide whether a pattern is contained or not in a sequence (*implicit choice of semantics*)
- The examples have been carefully selected to reveal the interpretation of one dimension of the semantics of NSP

Example of question: multiple occurrences

According to you, what are the sequences that contain the pattern $\boldsymbol{p} = \langle b\ \neg e\ f \rangle$?

id	Sequence
$\boldsymbol{o}_0$	$\langle b\ a\ f\ d\ b\ d\ f \rangle$
$\boldsymbol{o}_1$	$\langle b\ a\ f\ d\ e\ b\ d\ f \rangle$
$\boldsymbol{o}_2$	$\langle d\ b\ e\ c\ a\ d\ f\ b\ d\ e\ f \rangle$
$\boldsymbol{o}_3$	$\langle b\ a\ f\ b\ a\ e\ f \rangle$

- The user is invited to decide whether a pattern is contained or not in a sequence (*implicit choice of semantics*)
- The examples have been carefully selected to reveal the interpretation of one dimension of the semantics of NSP

- $\boldsymbol{o}_0$, $\boldsymbol{o}_1$ and $\boldsymbol{o}_3$ $\implies$ weak occurrence
- $\boldsymbol{o}_0$ $\implies$ strong occurrence
- $\boldsymbol{o}_1$ is a trap ... and is ignored
- other combination of ticks $\implies$ "other" semantics

Example of question: multiple occurrences

According to you, what are the sequences that contain the pattern $\boldsymbol{p} = \langle b\ \neg e\ f \rangle$?

id	Sequence
$\boldsymbol{o}_0$	$\langle b\ a\ f\ d\ b\ d\ f \rangle$
$\boldsymbol{o}_1$	$\langle b\ a\ f\ d\ e\ b\ d\ f \rangle$
$\boldsymbol{o}_2$	$\langle d\ b\ e\ c\ a\ d\ f\ b\ d\ e\ f \rangle$
$\boldsymbol{o}_3$	$\langle b\ a\ f\ b\ a\ e\ f \rangle$

- The user is invited to decide whether a pattern is contained or not in a sequence (*implicit choice of semantics*)
- The examples have been carefully selected to reveal the interpretation of one dimension of the semantics of NSP

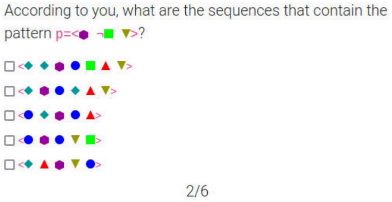
- $\boldsymbol{o}_0$, $\boldsymbol{o}_1$ and $\boldsymbol{o}_3$ $\Longrightarrow$ weak occurrence
- $\boldsymbol{o}_0$ $\Longrightarrow$ strong occurrence
- $\boldsymbol{o}_1$ is a trap ... and is ignored
- other combination of ticks $\Longrightarrow$ "other" semantics

Two alternative visualisations

According to you, what are the sequences that contain the pattern p=<b ¬a e>?

- ☐ <f f b d a c e>
- ☐ <f b d f c e>
- ☐ <d f b d c>
- ☐ <d b d e a>
- ☐ <f c b e d>

2/6



- Ease the use for unconfortable people with formal notations
- Prevent from being influenced by an implicit order on events [commented by some surveyed people]

→ we did not collect the information about who used which notation!

Outline

Gathering answers

Technical details

- survey in English
- hosted on a personal website (no specific tools used)

- 124 survey answers fully filled
  - 54 knowledgeable in data science
  - 27 knowledgeable in pattern mining
  - 23 knowledgeable in logic
  - 40 without specific knowledge in one of these two fields
  - 82 researchers

- Survey answers form a large tabular datasets (mainly boolean values)
- Analysis of answers with Formal Concept Analysis

  → Unsupervised identification of groups of people having the same kind of answers

Scope of the negation

Table: Result on the question about the scope of negation

Scope	Count	Percentage
Conform	101	81.4%
Conform except $\boldsymbol{s}_4$	9	7.3%
Alternative	14	11.3%

- $\langle f\ a\ c\ e\ b\rangle$ contains $\langle c\ \neg d\ e\rangle$? Possible different semantic from above

$$\neg e \Leftrightarrow \exists s_i \in \boldsymbol{s},\ i \in [\ldots],\ s_i \neq e$$

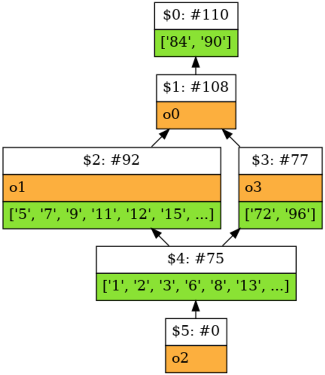
  $\rightarrow$ no such situation in the other questions!
- We keep the 110 valid answers in the remaining of the analysis

Occurrence dimension

Interpretation	Count	Percentage
Weak relation	75	69.2%
Strong relation	33	28.2%
Other	2	3.6%

→ 75 people in concept 3 (weak occurrences: o_0, o_1 and o_3)

→ 33 people in concept 1 (strong occurrences: o_0)



Conclusion

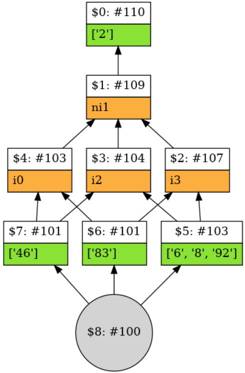
Their are two alternative interpretations in the panel: 70% weak / 30% strong occurrences.

Non-inclusion dimension

Interpretation	Count	Percentage
Partial non-inclusion	100	90.9%
Total non-inclusion	3	2.7%
Other	7	6.4%

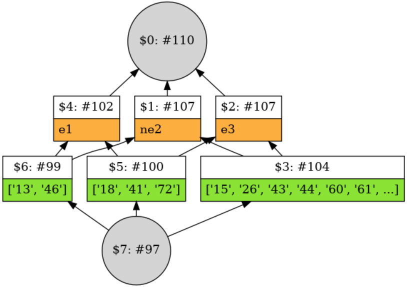
Conclusions

→ "Partial non-inclusion" seems to be the most intuitive notion for itemset non-inclusion.



Embedding dimension

Interpretation	Count	Percentag
Strict occurrence	97	88.2%
Soft occurrence	7	6.3%
Other	6	5.5%



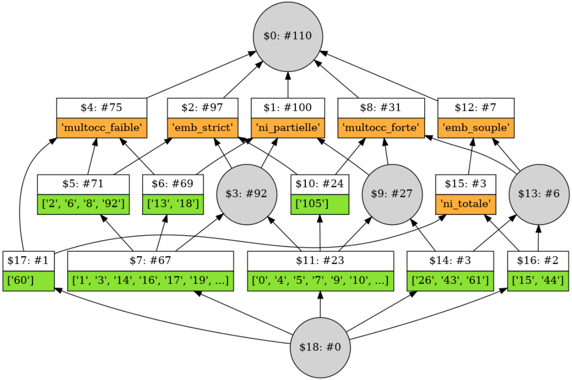
Conclusion

→ the "soft-occurrence" interpretation dominates

Global analysis

Conclusions: there are mainly two semantics that are intuitively used

- Partial non-inclu, soft embedding, strong containment at 23.9%
- Partial non-inclu, soft embedding, weak containment at 69.8%
- The other semantics are marginally represented



Results' conclusions and recommendations

Conclusions

- There are mainly two semantics that are intuitively used
- No statistical significant difference between the groups of people (with the characteristics we collected)
- None of the state of the art algorithms fits to the intuition, because of the partial non-inclusion

Recommendations

1. use only singletons in the negations. In this case, partial and total non-inclusions are equivalent
2. develop an alternative adapted to a partial interpretation of the non-inclusion
   → extend preferably NegPSpan regarding its management of multiple occurrences that meets the intuition of a larger number of people
3. promote the use of different syntaxes for each semantics

Discussion (about the methodology)

Known limits of the methodology

- Is the surveyed population representative of potential users of pattern mining algorithms?
  → not enough questions to describe the population!
- Non-redundancy of the questions:
  → strengthen the assignment of an interpretation by multiple questions per dimension
- "Small" number of answers:
  - it is not so small ... and the results are clear
  - people have conscientiously answered the questions (very poor rate of weird answers)
- Bias in the presentation of basic notions of sequential patterns
- Questionnaire is closely linked to the analysis framework proposed by Besnard and Guyet [BG20], more specifically:
  → syntactic restrictions
  → 18.5% did not answer as expected to the scope question!
  → Long interviews could complement these results

Outline

Conclusions

Our initial research questions

1. Are there "intuitive" semantics for patterns with negation?
   → There are two dominant ones!
2. Do the "intuitive" semantics correspond to those actually used by one of the algorithms?
   → No, because of the partial non-inclusion
3. What are the recommendations on the use of patterns with negations?
   → extend NegPSpan with partial non-inclusion
   → promote the use of different syntaxes for each semantics

Is pattern mining an "interpretable" data analysis technique?

- pattern mining outputs easy to present results, but
- the existing NSP mining algorithms may leads to data/pattern misinterpretation
- their interpretation requires additional information to prevent from misinterpretation

References I

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

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

Yongshun Gong, Chuanlu Liu, and Xiangjun Dong, *Research on typical algorithms in negative sequential pattern mining*, Open Automation and Control Systems Journal 7 (2015), 934–941.

Thomas Guyet and René Quiniou, *NegPSpan: efficient extraction of negative sequential patterns with embedding constraints*, Data Mining and Knowledge Discovery 34 (2020), no. 2, 563–609.

Sue-Chen Hsueh, Ming-Yen Lin, and Chien-Liang Chen, *Mining negative sequential patterns for e-commerce recommendations*, Proc. of Asia-Pacific Services Computing Conference, 2008, pp. 1213–1218.

Zhigang Zheng, Yanchang Zhao, Ziye Zuo, and Longbing Cao, *Negative-GSP: An efficient method for mining negative sequential patterns*, Proc. of the Australasian Data Mining Conference, 2009, pp. 63–67.