Discovering metric temporal constraint networks on temporal databases

Miguel R. Álvarez

Centro Singular de Investigación en Tecnoloxías da Información

Universidade de Santiago de Compostela

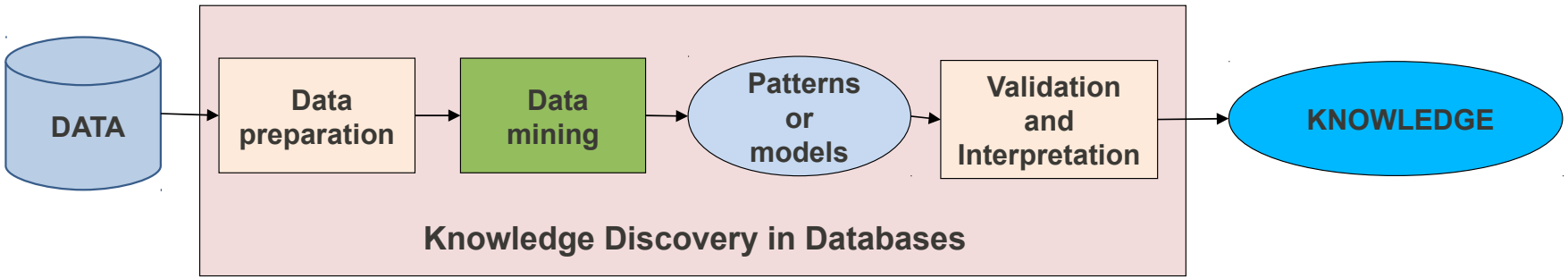
July 5, 2013

Contents

KDD

Motivation

A huge amount of data continuously collected in daily activities of all kinds of organisations.



KDD

To extract interesting, non trivial, implicit, **previously unknown** and potentially useful information from data.

KDD Data mining

A data mining process must be naturally **interactive** and **iterative**.

The integration of **background knowledge** about the processes under study is usually a requisite for the usefulness of the final result of the mining process.

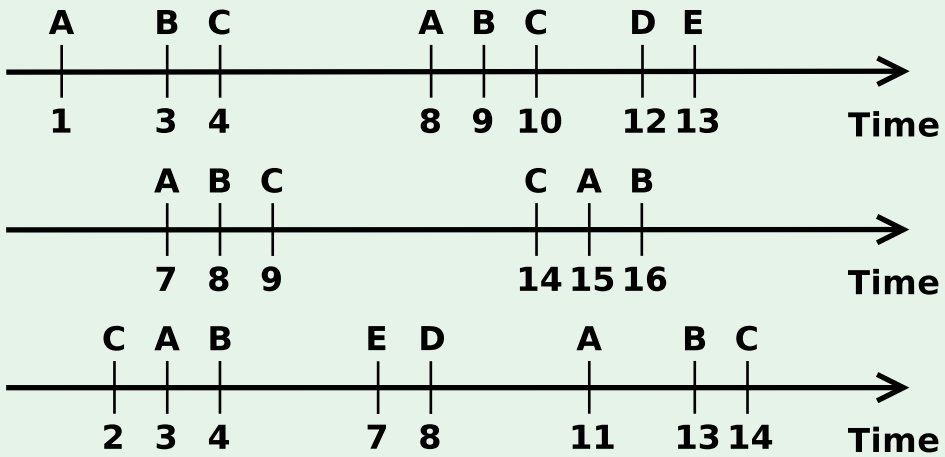
The more **expressive** the knowledge to be injected into the mining algorithms, the more complex the search on the database.

The development of **new efficient algorithms** that enable human experts to collaborate in the mining process in a more intuitive way is as an important challenge for future data mining proposals.

Sequential Data

In some application domains data consist of event sequences, where the order of the events is relevant to the interpretation of the underlying processes.

Sequences



Examples

- Importance of temporal relations in shopping habits.
- Importance of temporal relations of manifestations in diseases.

Sequence Data Mining

Objective

Find frequent sequences in a collection of sequential data.

A sequence is considered **frequent** if its presence in the data is higher than a user defined threshold.

Apriori strategy

Use of frequent sequences of size k to reduce the search scope of sequences of size $k + 1$.

Contents

Hypothesis and objectives

Hypothesis

Any effective temporal data mining technique should be based on some formalism for temporal reasoning and representation.

Objectives

Design and develop temporal data mining algorithms that improve the expressiveness of previous proposals.

- Discover **frequent temporal patterns** from a set of time-stamped event sequences.
- Represent patterns as metric temporal constraint networks.
- Consider **events** and **episodes** as temporal entities to be mined as part of a time-stamped event sequence.
- Allow a domain expert to introduce knowledge into the process.

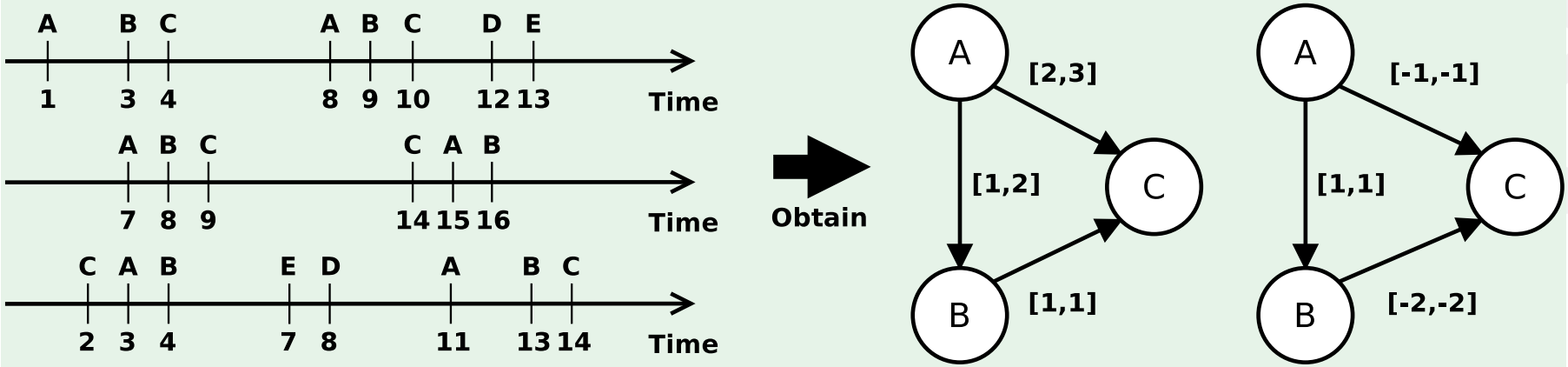
Contents

Stage 1: Algorithm description

Purpose

- To extract a set of frequent temporal patterns from a collection of event sequences.
- Each pattern represents a set of temporal arrangements between events considered sufficiently **similar** and **frequent**.

Intentions



Apriori strategy

Apriori-based algorithm

Searches for frequent patterns iteratively:

- Candidate generation.
- Frequency calculation.
- Non-frequent candidate removal.

STP patterns

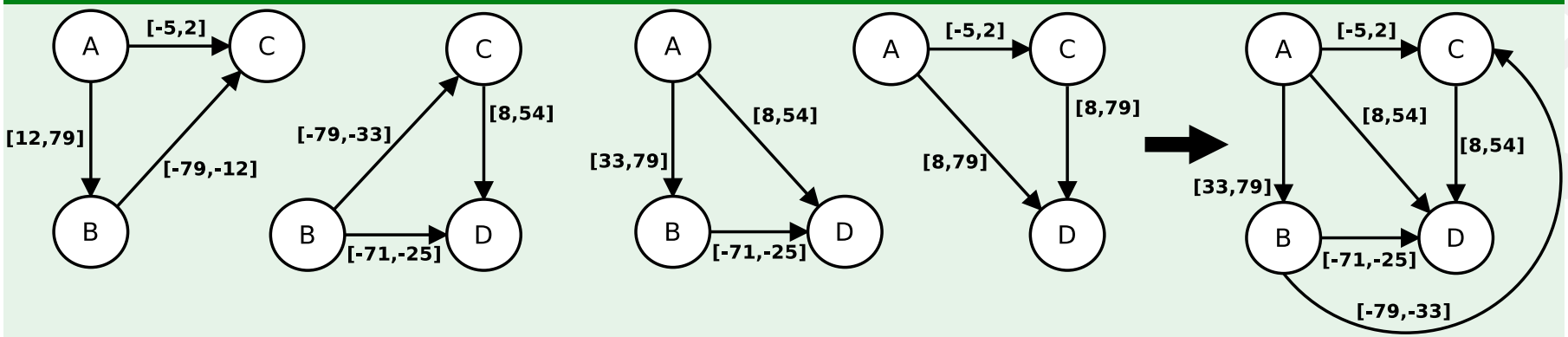
- Quantitative temporal constraint networks.
- Candidate patterns constructed from previous frequent patterns.
- Size two patterns obtained using a clustering procedure.

Algorithm steps

Candidate generation

- Each candidate of size k is built from the combination of k frequent patterns of size $k-1$.
- A *Floyd-Warshall* algorithm validates the consistency of the resulting network.

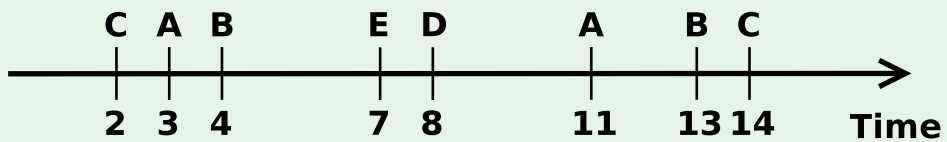
Combination example



Algorithm steps

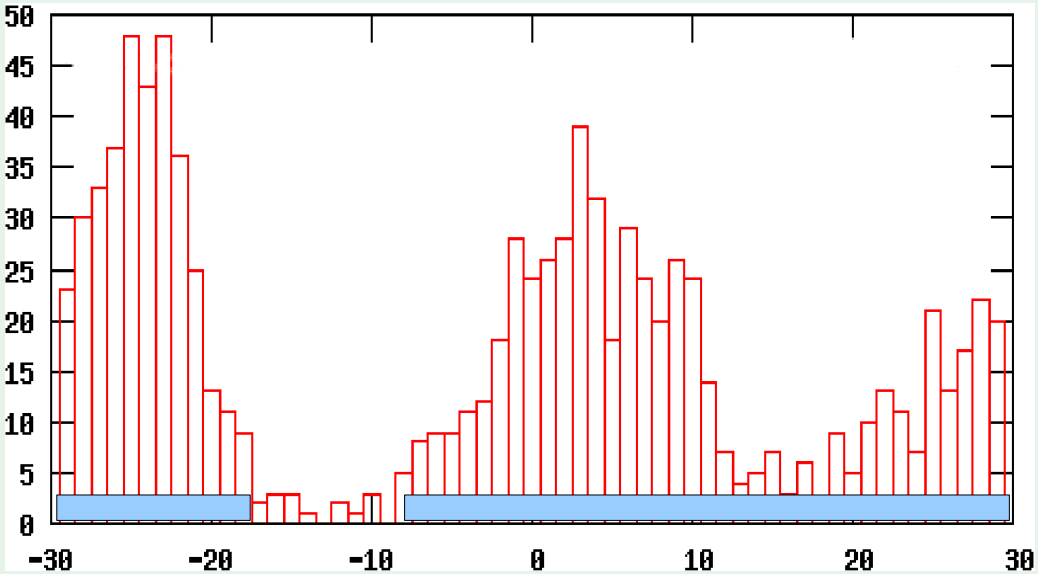
Size 2 frequency calculation

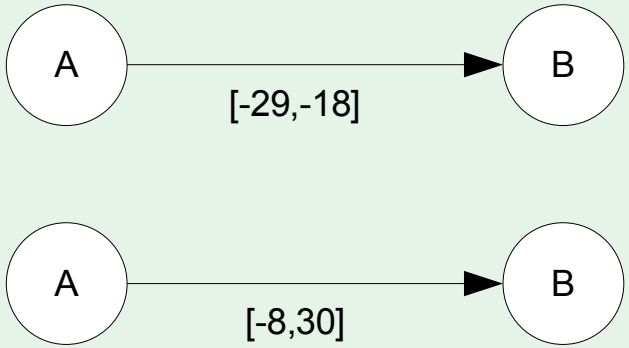
- Counts the number of occurrences of every temporal arrangement between every pair of event types found in the data collection.
- Builds a **temporal distance distribution** that undergoes a **clustering** procedure, obtaining a set of intervals.
- Each interval specifies **similar** and **frequent** temporal arrangements between the event types represented.



Algorithm steps

Clustering of temporal distance distributions





Clustering is only performed in iteration 2. Some temporal arrangements might be conditioned by the presence of other events.

Application domain: SAHS

Data set

Annotated database of 50 SAHS patients, consisting of 120,000 events from 8 event types and 280 hours of sleep.

Set of annotated event types

E_1 : *"start of airflow limitation"*

E_2 : *"end of airflow limitation"*

E_3 : *"start of abdominal movement limitation"*

E_4 : *"end of abdominal movement limitation"*

E_5 : *"start of thoracic movement limitation"*

E_6 : *"end of thoracic movement limitation"*

E_7 : *"start of O_2 desaturation"*

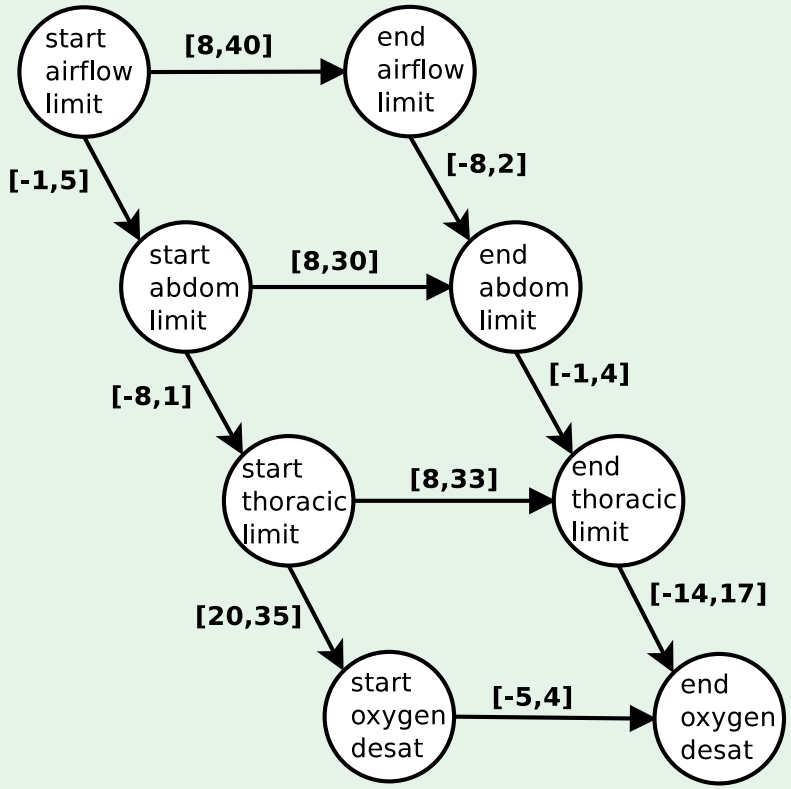
E_8 : *"end of O_2 desaturation"*

Application domain: SAHS

Basic algorithm

Results

Size	Frequent
1	8
2	46
3	165
4	303
5	297
6	160
7	46
8	6



Stage 2: Seed patterns

Objective

Provide the mining user mechanisms to introduce previous domain knowledge into the process in order to **constrain the search scope** and **refine the results**.

Knowledge provided

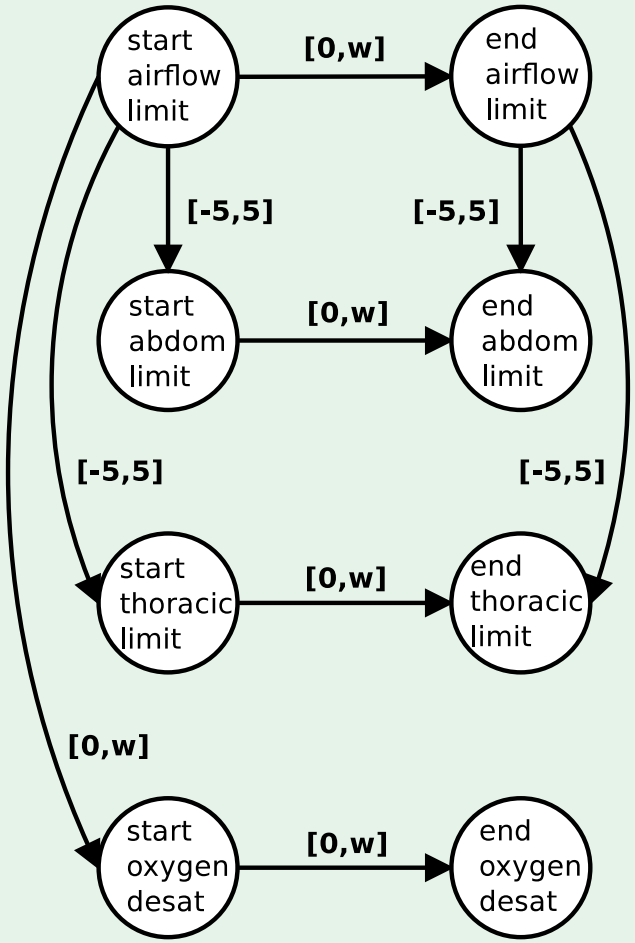
- The user specifies a set of event types of interest, and some constraints between them, constructing a **seed pattern**.
- All patterns involved in the mining process must be consistent with the knowledge provided.

Initialisation step

Find all frequent patterns of size two consistent with the seed pattern provided.

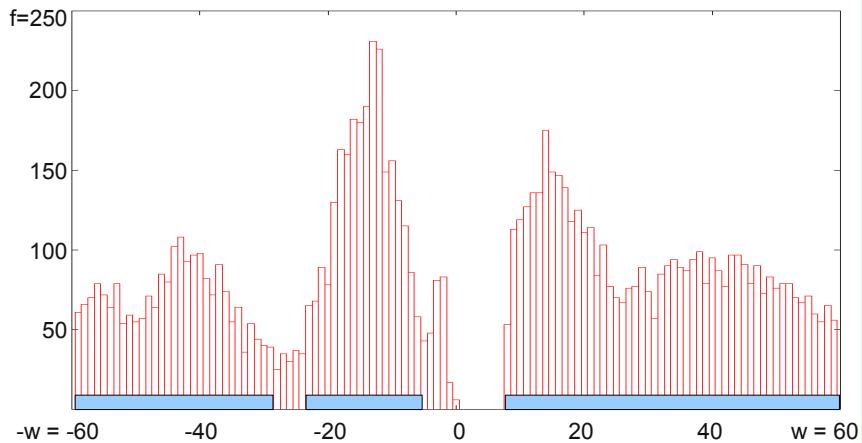
Seed patterns

Seed pattern example

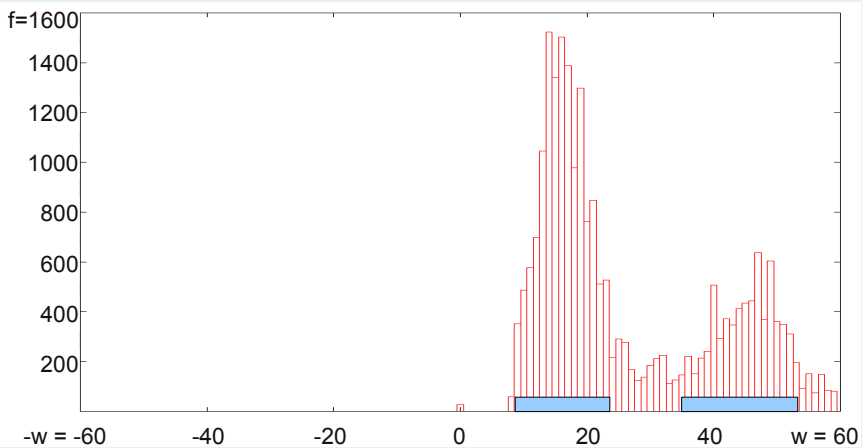


Application domain: SAHS
Using seed patterns

Without seed pattern



Using a seed pattern

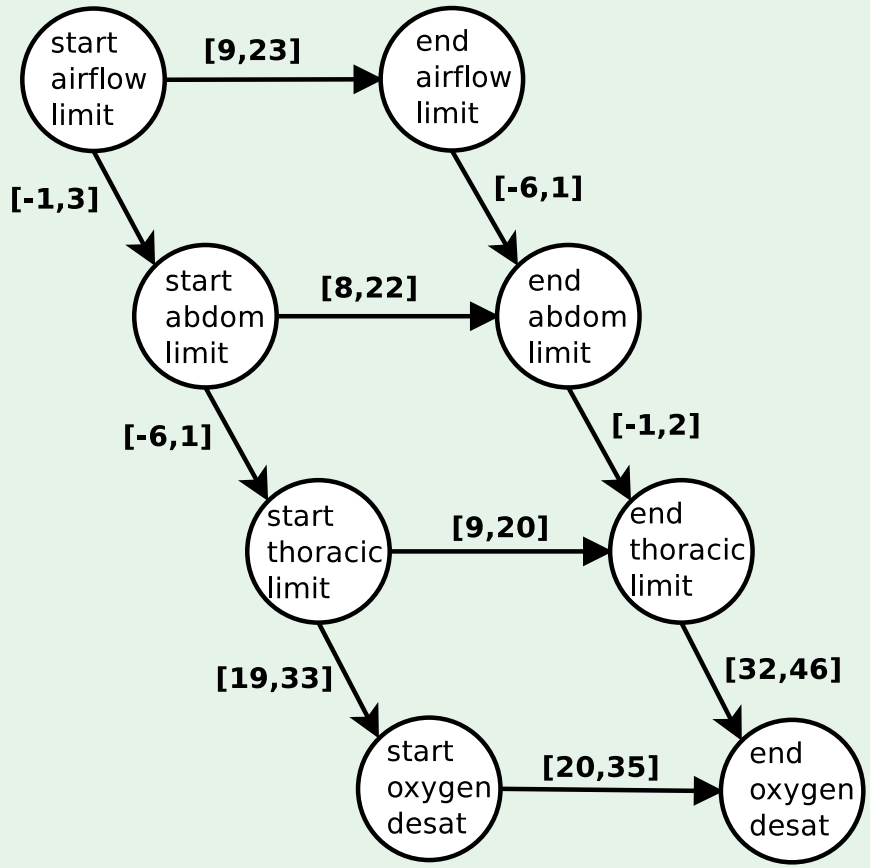


Application domain: SAHS

Using seed patterns

Results

Size	Frequent	
	seed	**w/o seed**
1	8	8
2	38	46
3	102	165
4	174	303
5	192	297
6	128	160
7	48	46
8	7	6
Total	689	1031



Stage 3: Episodes

Episodes

Events represent entities with no duration. Those entities with duration can be represented using the notion of **episode**, using two events to limit the beginning and ending of the episode.

Example

An airflow limitation episode can be represented by its corresponding beginning and ending events.

Constraints

- A beginning event can only be associated to one ending event.
- Both events must be present in the temporal window.

Application domain: SAHS

Defining episodes

Data set

Annotated database of 50 SAHS patients, consisting of 120,000 events from 8 event types and 280 hours of sleep.

Set of episode types involved:

G_1 : "*airflow limitation*"

G_2 : "*abdominal movement limitation*"

G_3 : "*thoracic movement limitation*"

G_4 : "*O_2 desaturation*".

Application domain: SAHS

Defining episodes

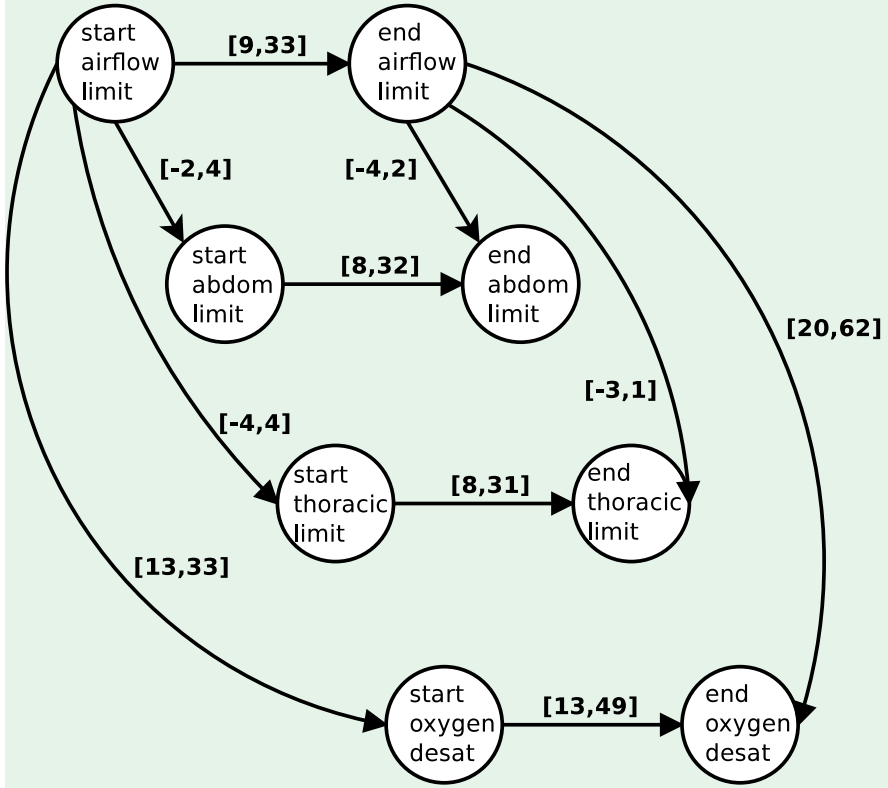
Results

Size	Frequent	
	seed	**w/o seed**
1	8	8
2	32	44
3	71	157
4	99	328
5	82	363
6	38	209
7	9	54
8	1	4
Total	340	1167

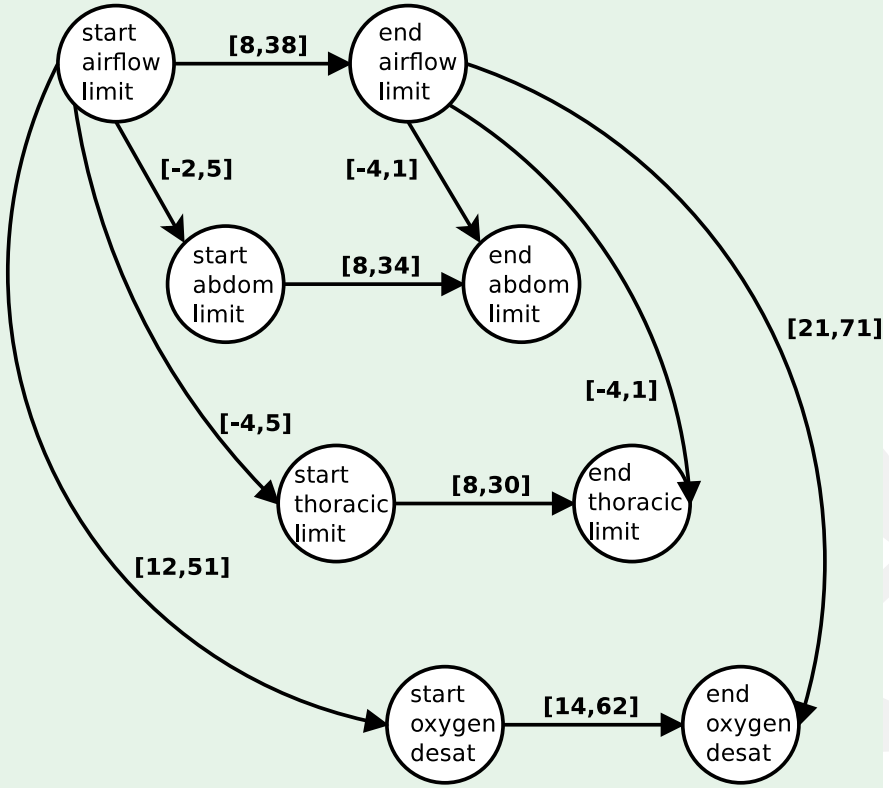
Application domain: SAHS

Defining episodes

Resulting patterns



a) Using seed pattern



b) Without seed pattern

Contents

Summary

Temporal data mining

The techniques developed are able to induce patterns represented as temporal constraint networks from collections of event sequences.

User interaction

- Users can participate in the mining process by introducing previous knowledge of the domain to focus the search.
- User provides knowledge as a seed pattern: a number of events of interest and some temporal constraints between them.
- This knowledge focuses the search process on those patterns that extend the seed, either by incorporating new event types or by refining the constraints in the seed pattern.
- Seed patterns improve the algorithm efficacy and efficiency:
  - The constraints specified by the user limit the patterns found to those consistent with the information provided.

Future work

Limitations

The STP formalism does not convey all the available information.

- Some temporal arrangements of the intervals are more frequent than others.
- Histograms are reduced to intervals.
- The shape of the temporal distance distribution is not used.
- Seed patterns lose linguistic nuances.

Proposals

Enhance or replace the STP representation.

- Associate density functions to temporal constraints.
- Use the FTCN representation, transforming the temporal distance distributions into possibility distributions.
- Extract and extend fuzzy seed patterns.

Future work

Limitations

The Apriori strategy produces a large amount of patterns.

- Costly candidate generation and frequency calculation steps.
- Increased memory requirements.

Proposal

Explore other search strategies, such as depth-first approaches.

Limitations

It is not possible to model the systematic absence of the occurrence of some event type during the occurrence of a pattern.

Proposal

Introduce negated event types that do not appear in the constraint network, yet represent that no occurrence of the event types may be found in the same temporal context of an occurrence of the pattern.

Thank you

Questions?