Efficient declarative sequence mining with ASP

Supervision:
- Thomas Guyet, IRISA/LACODAM (thomas.guyet@irisa.fr)

Context:
Sequence mining is a task that consists in extracting interesting subsequences from a collection of sequences. This pattern mining task is wide used to analyze behaviors from longitudinal traces collected on digital, physical or living systems. For instance, traces can be the logs of servers, it can be the purchase sequences of supermarket clients or the care pathways, ie the sequences of cares that sick people had.

The mining of sequences has been widely studied since two decades and very efficient algorithms have been designed. Nonetheless, this approach can not handle the complex nature of traces neither the expert knowledge of the data acquisition. This leads such algorithms to the pitfall of the pattern flood: they extract an incredibly large number of meaningless patterns which make tedious the data analyst task.

Some recent works proposed to use declarative paradigms (SAT, constraint programming – CP and logic programming) to implement sequence mining tasks and propose versatile tools that could integrate expert knowledge to refine the outputs.
The approach that is developed in Rennes is based on Answer Set Programming (ASP) which combines a high level language to formalize complex knowledge and also an efficient solver (*clingo*[1]) based on the most recent technologies of SAT and CP which are our competitors. This is done in closed relationship with the developers of the clingo solver in university of Potsdam.
Our previous research line was to develop pure ASP encoding of the sequence mining task in order to illustrat, in the field of care pathway analysis, that ASP is an interesting solution to address the design of new complex mining task without knowledge of mining algorithms, but it lacks of efficiency to address large databases.

The objective of this internship is to explore the hybrid-solving capabilities of *clingo* to increase the efficiency (in time and in memory) of ASP approaches. Hybrid solving enables to embed in the solver some procedural codes, called propagators. The propagators are used, the less versatile would be the programs. And the question lies in the definition of propagators that makes programs more efficient while keeping a certain generality.

The main steps of this work will be:
- the study of the state of the art of declarative and hybrid sequence mining
- to get familiar with the *clingo* solver
- to propose and evaluate different propagators within the *clingo* solver
- the compare this approach with competitors in CP and SAT
- the implement *smart* mining task to illustrate the versatility of the proposal

References
- Gerbser, Martin, Guyet, Thomas, Quiniou, René, Romero, Javier, Schaub and Torsten. "Knowledge-based Sequence Mining with ASP". Proceedings of IJCAI : p. to appear. 2016

1 https://potassco.org/clingo/

- Aoga, J. O., Guns, T., & Schaus, P. (2016, September). An efficient algorithm for mining frequent sequence with constraint programming. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 315-330).
- Negrevergne, B., & Guns, T. (2015, May). Constraint-based sequence mining using constraint programming. In *International Conference on AI and OR Techniques in Constriant Programming for Combinatorial Optimization Problems* (pp. 288-305)
- Paramonov, S., Stepanova, D., & Miettinen, P. (2017, July). Hybrid ASP-Based Approach to Pattern Mining. In *International Joint Conference on Rules and Reasoning* (pp. 199-214).
- Thomas Guyet, Yann Dauxais and André Happe. "Declarative sequential pattern mining of care pathways". Proceedings of Conference on AIME : p. 261–266. 2017
- Guyet, T., Moinard, Y., Quiniou, R., & Schaub, T. (2018). Efficiency Analysis of ASP Encodings for Sequential Pattern Mining Tasks. In *Advances in Knowledge Discovery and Management* (pp. 41-81).

Expected profile and skills

- Logic programming/constraint programming notion
- Python or C++ for propagators implementation
- Interest in data science
- Curiosity
- Scientific English