MDL for Robust Periodic Pattern Mining
Or How to Detect Habit Changes ?

Peggy Cellier, IRISA - SemLIS, INSA Rennes, peggy.cellier@irisa.fr
Esther Galbrun, Aalto University, esther.galbrun@aalto.fi
Alexandre Termier, IRISA - Lacodam, Univ. de Rennes 1 Alexandre.Termier@irisa.fr

Location: IRISA, team Lacodam/SemLIS, Rennes

Keywords: data science, data mining, pattern mining, Minimum Description Length (MDL)

Context and description of the intership:
Event logs are among the most ubiquitous types of data nowadays. They can be machine generated (server logs, database transactions, sensor data) or human generated (ranging from hospital records to life tracking, a.k.a. quantified self), and are bound to become ever more voluminous and diverse with the increasing digitisation of our lives and the advent of the Internet of Things (IoT). Such logs are often the most readily available sources of information on a system or process of interest. It is thus critical to have effective and efficient means to analyse them and extract the information they contain.

To help understand the characteristics of the underlying recurrent phenomena recorded in logs, some approaches such as periodic pattern mining [1] have been proposed. Those algorithms allow to discover periodic repetitions of sets or sequences of events amidst unrelated events. However, such algorithms suffer from the traditional plague of pattern mining algorithms: they output too many patterns (up to several millions), even when relying on condensed representations [2].

In [3], a novel approach for mining periodic patterns using a MDL criterion has been proposed. The Minimal Description Length (MDL) principle [8] is a concept from information theory based on the insight that any structure in the data can be exploited to compress the data, and aiming to strike a balance between the complexity of the model and its ability to describe the data. This principle has been adapted to pattern set mining and has given rise to a fruitful line of work [4,5,6,7]. In [3] an expressive pattern language has been defined as well as the associated encoding scheme which allows to compute a MDL-based score for a given periodic pattern collection and a sequence. With this approach complex period pattern with nested cycles can be found, as for instance :

> « Starting Monday at $7$:$30$ AM,
> **wake up**, then, $10$ minutes later, **prepare coffee**,
> repeat every $24$ hours for $5$ days,
> repeat this every $7$ days for $3$ months »

The aim of the internship is to building up on the existing work [3] to make the approach more robust to variations, such as gaps in the patterns or concept drift. Some experiments will be conducted on datasets from human logs and process logs. The expected research work requires a taste for theory, algorithmic and experiments.

[1] B. Özden, S. Ramaswamy, and A. Silberschats. **Cyclic association rules**. ICDE, 1998.

[2] P. Lopez-Cueva, A. Bertaux, A. Termier, J .-F. Méhaut, and M. Santana. **Debugging embedded multimedia application traces through periodic pattern mining**. Int. Conf. On Embedded Software, 2012.

[3] Esther Galbrun, Peggy Cellier, Nikolaj Tatti, Alexandre Termier, and Bruno Crémilleux. **Mining Periodic Patterns with a MDL Criterion**. ECML-PKDD 2018.

[4] J. Vreeken, M. van Leeuwen, and A. Siebes. **Krimp : Mining itemsets that compress**. DMKD, 2011.

[5] F. Bonchi, M. van Leeuwen, and A. Ukkonen. **Characterizing uncertain data using compression**. SDM, 2011.

[6] N. Tatti and J. Vreeken. **The long and the short of it : Summarising event sequences with serial episodes**. KDD, 2012.

[7] A. Bhattacharyya and J. Vreeken. **Efficiently summarising event sequences with rich interleaving patterns**. SDM 2017.

[8] P. Grünwald. **Model Selection Based on Minimum Description Length**. Journal of Mathematical Psychology, 2000.