IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

POLYTECHNIQUE MONTRÉAL
TECHNOLOGICAL UNIVERSITY

PhD Position:

Counting and sampling of solutions for anytime pattern discovery

1 - Context and funding

Constraint-based pattern mining is a fundamental data mining task, extracting locally interesting patterns to be either interpreted directly by domain experts, or to be used as descriptors in downstream tasks, such as classification or clustering. Recently, this approach has been challenged by an increasing focus on user-centered, interactive, and anytime pattern mining [2]. This new paradigm stresses that users should be presented quickly with patterns likely to be interesting to them, and typically affect later iterations of the mining process by giving feedback. A powerful framework for taking a variety of user feedback into account is pattern mining via constraint programming (CP). Much of the current focus in this domain is on user-centered/interactive mining, particularly the ability to elicit and exploit user feedback [1,2]. An important aspect of requesting such feedback is that the user be quickly presented with diverse results. If patterns are too similar to each other, deciding which one to prefer can become challenging, and if they appear in several successive iterations, it eventually becomes a slog. Similarly, a method that produces diverse results but takes a long time to do so, risks that the user checks out of the process. Sampling algorithms [3,4] can circumvent these negative complexity results by sampling a representative set of patterns, according to a probability distribution that is proportional to a given quality measure, without explicitly enumerating all patterns. These approaches can substantially improve efficiency as well as controllability of pattern discovery processes.

While a number of pattern sampling approaches have been developed over the past years, they are either inflexible (as they only support a limited number of quality measures and constraints), or do not provide theoretical guarantees concerning the sampling accuracy. At the algorithmic lever, they mainly rely on the Markov Chain Monte Carlo random walks over the pattern space [3,4], or a special purpose sampling procedure tailored for a restricted set of itemset mining tasks [8]. Other approaches use recent advances in sampling solutions in SAT to partition the search space into cells using random XOR constraints and then extracts a pattern from a randomly selected cell [5]. This solution space reduction approach has also been transposed to constraint programming [6,7].

In the last decade, data mining has been combined with constraint programming to model various data mining problems [9,10,11]. The main advantage of CP for pattern mining is its declarativity and flexibility, which include the ability to incorporate new user-specified constraints without the need to modify the underlying system. Similarly, some recent works in CP for counting solutions of individual constraints have been proposed [12,13,14]; in particular Truchet and Pesant collaborated on this subject [12]. All these methods to count or to sample rely on close mathematical techniques or models, as has been shown in the case of SAT problem [15].

2 - Research project

The focus of this thesis is to **develop new methods for sampling and counting for interactive, and anytime pattern discovery**. The methods will be studied through the prism of new class of constraints, pattern constraints, dedicated to model some complex tasks in data mining. These constraints, which are based on new structures, remain a scientific challenge. Thanks to the flexibility of the CP framework, a variety of pattern quality measures will be considered to sample patterns while still providing strong theoretical guarantees.

3 -Team supervision and PhD registration

The university partner Polytechnique Montréal, and more specifically the Quosséça research center, is a major actor in AI in Canada and at the international. It is involved in a large number of industrial and academic projects.

Supervisors:

- Gilles Pesant, laboratoire Quosséça, Polytechnique Montréal, gilles.pesant@polymtl.ca
- Samir Loudni, IMT Atlantique, LS2N, samir.loudni@imt-atlantique.fr
- Charlotte Truchet, Nantes University, LS2N, charlotte.truchet@univ-nantes.fr

Gilles Pesant and Charlote Truchet have recently collaborated as part of Giovanni Lo Bianco's thesis, on the enumeration of solutions for the global cardinality constraint. This thesis will extend this work to the context of interactive pattern discovery. Samir Loudni brings his expertise on the triptych (A) constraints, (B) symbolic data mining, (C) preferences.

The student will divide its time into two periods, one in Canada and one in France, where frequent working visits and collaborations will take place from one institution to the other.

4 - Candidate profile

The successful candidate will have (or will soon obtain) an MSc (or similar) in Computer Science or related subject.

Required skills:

- Constraint Programming, mathematics, machine learning.
- Strong facility in software engineering and implementation (Java, Python)
- Strong mathematical and formal foundations
- A good command of written and oral English

5 - How to apply

Send a leler of motivation, transcript of grades, and your CV to Pr Samir Loudni (samir.loudni@imt-atlantique.fr) and Pr Gilles Pesant (gilles.pesant@polymtl.ca) with the subject beginning with [AI PhD].

6 – References

[1] Dzyuba, V., Leeuwen, M. v., Nijssen, S. et De Raedt, L. (2014). Interactive learning of pattern rankings. Int. Journal on Artificial Intelligence Tools,23(06):32 pages.

[2] van Leeuwen, M. Interactive Data Exploration Using Pattern Mining. Interactive Knowledge Discovery and Data Mining in Biomedical Informatics 2014: 169-182.

[3] Al Hasan, M. et Zaki, M. J. (2009). Output space sampling for graph patterns. Proc. of the VLDB Endowment, 2(1):730–741.