Pattern Sampling Under User-Constraints on Heterogeneous Data

PhD thesis Offer in Computer Science

Year 2025

Location:

- GREYC Laboratory, CNRS UMR 6072, University of Caen, 14000, Caen. With regular interactions with the Constraints and Learning team at LIFO Laboratory, EA 4022 – University of Orléans.

Scientific Context of the Internship

This PhD thesis is part of the FIDD project (Facilitated Exploration: Interactive Constraint-Driven Data Mining) funded by the ANR (French National Research Agency), which is planned to start in October 2025. The main objective of the FIDD project is to improve the user experience in the interactive data mining loop, as shown in Figure 1, by leveraging constraints to capture user interests and guide the data mining process effectively. An application is considered to improve patient care by and between radiologists based on their interactions and behaviors.

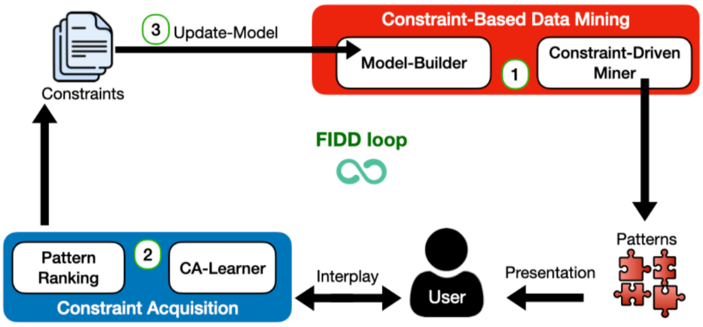

Figure 1: The FIDD loop with step 1: declarative data mining (modeling/solving); step 2: Constraint acquisition in data mining; step 3: Model adaptation/evolution.

Scientific Background

Pattern mining [1] involves extracting recurring patterns or models from a dataset to generate meaningful knowledge. However, to reduce processing time and give users more control, the late 2000s/early 2010s saw the development of *interactive mining* methods [7]: at each iteration, a small set of patterns is presented to the user, who examines these partial results and provides feedback that the algorithm incorporates in subsequent iterations. Due to the large number of extracted patterns, such an approach requires pattern sampling techniques, as proposed in [4, 5, 6, 3], to select a representative subset of patterns. These techniques reduce computational time complexity and facilitate analysis while preserving the essence of the information contained in the database patterns.

In these techniques, pattern sampling is often performed proportionally to a measure reflecting user interest or by incorporating constraints to restrict the search space. Thus, the sampling process can integrate constraints to influence the sampling itself or target specific patterns that satisfy certain defined properties. More formally, the sampling problem considered in this thesis is formulated as follows [4, 2]: given a database S, a pattern language $\mathcal{L}$, a set of constraints C, and a quality measure $\phi : \mathcal{L} \rightarrow \mathcal{R}$, randomly sample patterns that satisfy the constraints in C with a probability proportional to their quality.

Objectives

This PhD thesis focuses on the development of advanced pattern sampling techniques to support interactive data exploration, with particular emphasis on heterogeneous data and user-centered analysis. It is structured around two core research objectives:

1. Flexible Pattern Sampling for Heterogeneous Data: The first objective is to develop generalizable and scalable pattern sampling methods that can operate effectively across diverse data types, including binary, numerical, and sequential data. This entails addressing several key challenges, such as creating unified data representations, selecting appropriate interestingness measures, and ensuring efficient sampling despite the complexity of the data. We plan to leverage declarative frameworks such as constraint programming, SAT solving, or linear programming to handle the constraints for the sampling process. These methods will be embedded within interactive analysis workflows to support iterative exploration, enabling human-in-the-loop pattern discovery and continuous refinement of insights.

2. User-Guided Sampling and Adaptation to Evolving Constraints: The second objective is to incorporate user feedback and constraints directly into the sampling process. This reflects the dynamic nature of real-world data exploration, where users iteratively refine their queries, focus areas, and interpretations over time. The work will involve designing algorithms that can adapt to these evolving constraints, manage uncertainty, and personalize the exploration process. This includes learning from implicit and explicit feedback and adjusting sampling strategies to reflect changing user intent and data context.

This thesis offers opportunities to contribute at the intersection of machine learning, pattern mining, constraint reasoning and human-computer interaction. It is especially suitable for candidates interested in developing intelligent, adaptive systems that facilitate meaningful, user-driven data exploration.

Steps and Deliverables

The PhD project will proceed through the following major steps, each aligned with concrete deliverables:

1. Design and Implementation of a Flexible Pattern Sampler for Heterogeneous Data: Develop a unified framework that supports pattern sampling across various data types (binary, numerical, sequential) by unifying constraint specifications and pattern languages. This sampler will serve as a foundation for building adaptable and extensible interactive mining systems.

2. Development of Mechanisms to Detect and Respond to Evolving User Needs and Data Dynamics: Propose and implement strategies to monitor changes in user feedback, interests, or underlying data, enabling the generation of continuously adaptive models. This will support sustained, context-aware pattern discovery in interactive environments.

3. Evaluation Protocols for Interactive Pattern Mining: Design and conduct rigorous testing methodologies to assess the effectiveness, responsiveness, and usability of the proposed interactive systems. This includes both quantitative performance metrics and user-centered evaluation to validate system adaptability and insight facilitation.

Keywords:

Sampling methods, Pattern mining, Constraint Programming, Machine Learning.

Salary:

Approximately 2300 € gross.

Supervision Team:

- Abdelkader Ouali (GREYC – University of Caen).
- Thi-Bich-Hanh Dao (LIFO – University of Orléans).
- Bruno Crémilleux (GREYC – University of Caen).
- Christel Vrain (LIFO – University of Orléans).
- Albrecht Zimmermann (GREYC – University of Caen).

Desired Profile

Master's level (or equivalent) in computer science (or applied mathematics) with an interest in artificial intelligence, constraint programming, and data mining. Programming skills mainly in Java and Python as well as a good understanding of data mining algorithms and constraint solving will be appreciated. The working language is French or English.

How to Apply

Interested candidates are invited to submit the following documents, in PDF format only, to both of the following email addresses: abdelkader.ouali@unicaen.fr and thi-bich-hanh.dao@univ-orleans.fr:

- A curriculum vitae (CV)
- A cover letter detailing your qualifications, relevant experiences, and motivation for applying to this PhD project;
- Academic transcripts from your Bachelor's and Master's degrees (or equivalent qualifications for engineering schools);
- If available, names and contact information for academic or professional references (e.g., professors or supervisors familiar with your work);
- A link to your personal project repositories (e.g., GitHub), showcasing any relevant technical or research projects.

References

[1] Charu C. Aggarwal and Jiawei Han, editors. *Frequent Pattern Mining.* Springer, 2014.

[2] Arnaud Soulet Arnaud Giacometti. Dense neighborhood pattern sampling in numerical data. In *(SDM)*, San Diego, California, 2018.

[3] Anes Bendimerad, Jefrey Lijffijt, Marc Plantevit, Céline Robardet, and Tijl de Bie. Gibbs Sampling Subjectively Interesting Tiles. In *Advances in Intelligent Data Analysis {XVIII} - 18th International Symposium on Intelligent Data Analysis (IDA 2020)*, Konstanz (on line), Germany, April 2020.

[4] Mario Boley, Claudio Lucchese, Daniel Paurat, and Thomas Gärtner. Direct local pattern sampling by efficient two-step random procedures. In Chid Apté, Joydeep Ghosh, and Padhraic Smyth, editors, *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 582–590. ACM, 2011.

[5] Mario Boley, Sandy Moens, and Thomas Gärtner. Linear space direct pattern sampling using coupling from the past. In Qiang Yang, Deepak Agarwal, and Jian Pei, editors, *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 69–77. ACM, 2012.

[6] Mohammad Al Hasan and Mohammed J. Zaki. Musk: Uniform sampling of k maximal patterns. In *SDM*, 2009.

[7] Matthijs Van Leeuwen. Interactive data exploration using pattern mining. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, pages 169–182. Springer, 2014.