Master 2 Research Internship
Incremental Clustering with Declarative Methods

This master internship is part of a national project HERELLES, supported by ANR (Agence Nationale de la Recherche), starting on November 2020. In this project we will have a PhD funding, which should start in 2021. The internship is likely to lead to the PhD thesis.

Clustering is an important task in Data Mining, which aims at partitioning data instances into groups to find the underlying structure of the data. Clustering has been extended to constrained clustering, which allows to integrate prior knowledge, in order to make clustering task more accurate. Prior knowledge are integrated in the form of constraints. Most constrained clustering methods request the specification of all the constraints before the subsequent running of the methods. In many applications, it is more reasonable to allow the user to inject new information in the form of constraints on a clustering result. Constraints can be pairwise must-link or cannot-link constraints, which state that two instances must be or cannot be in the same cluster, or can be constraints on the clusters, stating bounds on their size or their diameter, or can be operations on clusters, such as split a cluster or merge two clusters, etc.. The constrained clustering process therefore becomes incremental. In this incremental setting, it is essential to take advantage of the information given by the user to make improvements on the solution. At the same time, in order to avoid confusing the user, new clustering should not be too different from the previous one.

We will consider declarative methods (Constraint Programming [1,2], Integer Linear Programming [3]) which offer the expressiveness and constraint satisfaction. In this internship, we aim at:

1. Developing a mechanism that eases the integration of feedback on a given clustering.
2. Identifying important constraints in order to take advantage of the information given by the user. This could be done by determining or improving a measure on the utility of the constraints [4].
3. At the same time, limiting the perturbation of the new clustering compared to the previous one. A measure of clustering similarity need to be defined, which can be either statistic or more explanatory.

Required skills:

- Experience in machine learning, data mining, computer programming or applied mathematics is highly appreciated.
- French and/or English are the working languages.

Candidates are encouraged to contact us as soon as possible. Start is expected on mid-January 2021. The duration will be 5 or 6 months. The complete application consists of the documents below, which should be sent as a single PDF file to Thi-Bich-Hanh Dao (thi-bich-hanh.dao@univ-orleans.fr):

- CV
- One-page cover letter (clearly indicating available starting date as well as relevant qualifications, experience and motivation)
- University certificates and transcripts (both B.Sc and M.Sc degrees marks)
- Contact details of up to three referees
- Possibly an English language certificate and a list of publications
- Attention: all documents should be in English or in French.

Supervisors: This master internship will take place in Orléans and will be supervised by members of the ANR Project Herelles in both laboratories LIFO in Orléans and GREYC in Caen. The persons involved in the supervision are:

- LIFO: Thi-Bich-Hanh Dao, Marta Soare, Christel Vrain
- GREYC: Samir Loudni, Abdelkader Ouali

References:

[1] Thi-Bich-Hanh Dao, Khanh-Chuong Duong, Christel Vrain: Constrained clustering by constraint programming. Artificial Intelligence 244: 70-94 (2017)
[2] Chia-Tung Kuo, S. S. Ravi, Thi-Bich-Hanh Dao, Christel Vrain, Ian Davidson: A Framework for Minimal Clustering Modification via Constraint Programming. AAAI 2017: 1389-1395
[3] Abdelkader Ouali, Samir Loudni, Yahia Lebbah, Patrice Boizumault, Albrecht Zimmermann, Lakhdar Loukil: Efficiently Finding Conceptual Clustering Models with Integer Linear Programming. IJCAI 2016: 647-654
[4] Ian Davidson, Kiri Wagstaff, Sugato Basu: Measuring Constraint-Set Utility for Partitional Clustering Algorithms. PKDD 2006: 115-126