Offre de thèse

Titre de la thèse :

Découverte de règles associant données hétérogènes (images et textes) et retours utilisateurs dans le cadre du clustering collaboratif.

Contexte et but de la thèse :

Cette thèse s'inscrit dans le cadre du projet Herelles financé par l'ANR (Agence Nationale de la Recherche), projet qui a démarré en novembre 2020. Herelles a pour but de définir un cadre théorique et opérationnel sur le clustering collaboratif avec une application phare portant sur des séries temporelles venant d'images de télédétection.

Dans ce cadre, une boucle d'interaction est mise en place entre les méthodes de clustering collaboratif et l'utilisateur afin que celui-ci puisse intervenir en ajoutant des contraintes sur les résultats des clusterings ; ces contraintes sont prises en compte dans les clusterings produits dans les itérations suivantes (cf Fig. 1). Cette démarche permet à l'utilisateur de converger plus rapidement sur la découverte d'une information pertinente.

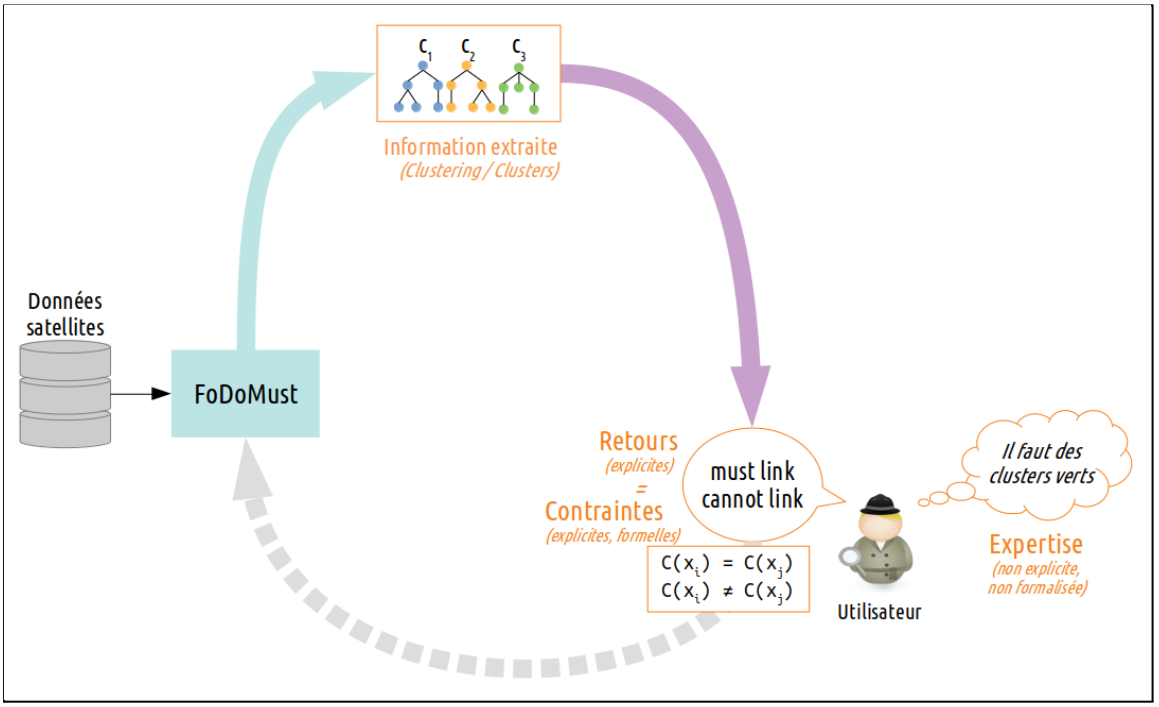

Fig. 1 : Le processus d'Herelles à l'heure actuelle.

Cependant, l'utilisateur ne sait pas toujours formuler les contraintes exprimant son intérêt. Le but de cette thèse est de concevoir de nouvelles méthodes pour l'aider (i) à représenter les données en combinant des informations issues des images et textes, (ii) à prendre en compte les retours utilisateurs pour découvrir des règles explicitant l'intérêt de l'utilisateur.

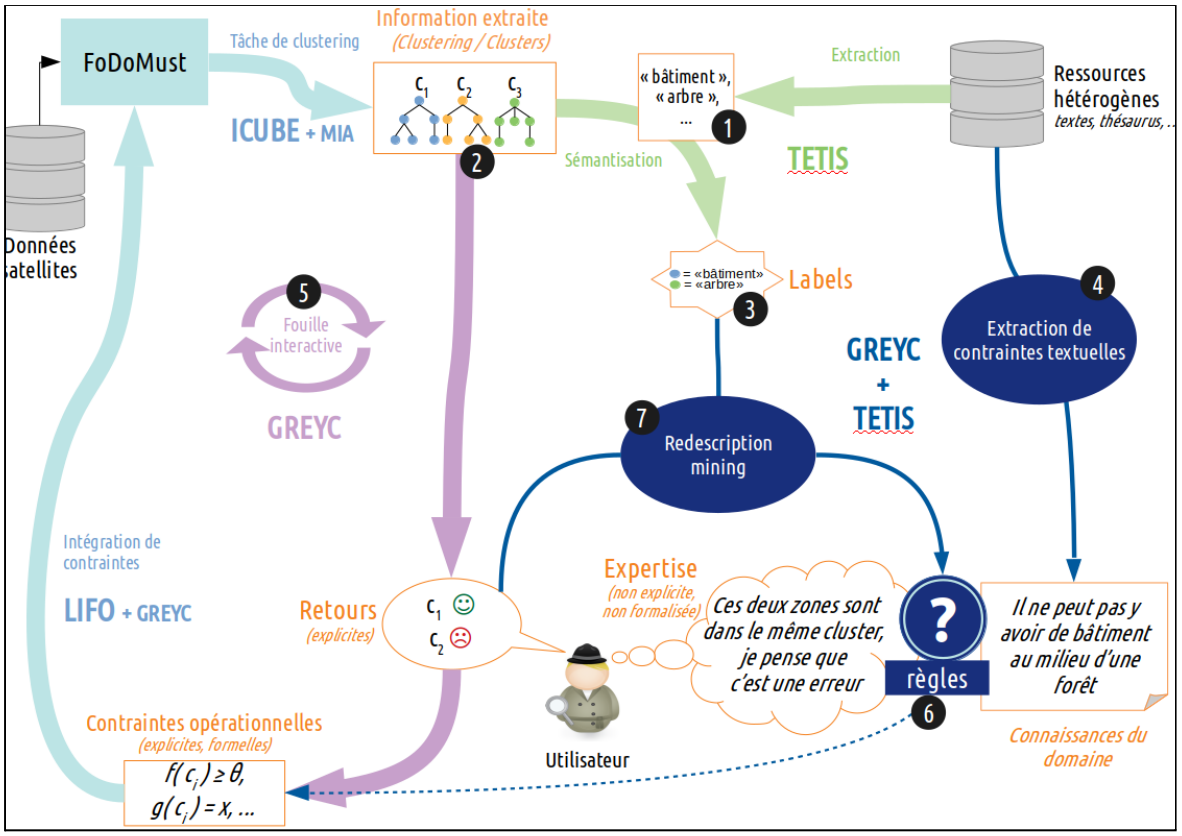

Fig. 2 : Le processus général de la thèse (en bleu) et son intégration au sein d'Herelles.

Description du sujet :

Dans le cadre de ce projet de thèse, une représentation des objets (sols, bâtiments,...) sera proposée ; elle reposera sur une combinaison de descripteurs (issus des images et textes) permettant de "sémantiser" les clusters (labellisation des clusters, cf. ❶ sur le schéma Fig. 2). Par exemple, avec une image on repère des *bâtiments* ou une *végétation*, avec des textes, leurs *fonctions* ou leur *type*[1] et on en déduit qu'un cluster représente des *bâtiments industriels ou des arbres*. L'ensemble de ces descripteurs formera une représentation originale et "sémantisée" qui associe une information calculée à partir des images satellites (compacité, spectre NDVI, nombre de points par clusters, etc.) *(❷)* et une information issue des ressources textuelles liée aux images (par exemple, mots-clés apportant une dimension thématique, troponymes apportant une dimension spatiale) (❸). En enrichissant la description des clusters, cette sémantisation va aider les utilisateurs dans l'analyse d'un clustering et de vérifier leur cohérence par rapport aux contraintes textuelles (❹) qui

[1] ou de manière plus précise leur espèce commes les *troènes du Japon* placés sur la place de la Comédie à Montpellier : https://www.midilibre.fr/2020/10/05/montpellier-les-arbres-de-la-comedie-font-reagir-les-habitants-9119110.php

peuvent être extraites de documents dédiés[2] (par exemple, *"l'ouverture à l'urbanisation de la zone X de Montpellier est subordonnée à la desserte par des transports collectifs"*).

D'autre part, ces clusters sémantiquement enrichis vont permettre de générer des règles de nature différente, intégrant à la fois cette sémantique et les retours de l'utilisateur. Ces retours sont obtenus de manière itérative, en s'appuyant sur le principe de la fouille interactive (❺). Un aspect remarquable de cette démarche est que celle-ci permet d'apprendre, à partir des retours de l'utilisateur, son intérêt modélisable sous forme de règles (❻). Ce résultat s'explique par le fait qu'un utilisateur peut être capable d'exprimer ce qui lui convient (ou pas) dans l'information extraite alors qu'il ne sait pas formaliser lui-même cet intérêt. Dans cette thèse, ce principe sera mis en œuvre via le redescription mining (❼). Le redescription mining propose "automatiquement" des règles intégrant différentes vues sur des objets. Ces vues sont ici l'information issue des images satellites, celle issue des ressources textuelles et les retours utilisateurs. L'originalité sera de combiner dans un même formalisme cette information hétérogène. Le but est de rendre possible la découverte de règles comme "l'utilisateur recherche des grandes zones végétales -clusters avec des points éloignés et ayant un indice NDVI élevé- et qui sont aussi des zones urbanisables". Pour cela, il est nécessaire de prendre en compte le fait que l'une des vues d'une redescription capture les retours utilisateurs, ce qui nécessite le développement de nouvelles méthodes de redescription mining.

Références :

- Jacques Fize, Mathieu Roche, Maguelonne Teisseire: Could spatial features help the matching of textual data? Intell. Data Anal. 24(5): 1043-1064 (2020)
- Esther Galbrun, Pauli Miettinen. Redescription Mining: An Overview. IEEE Intelligent Informatics Bulletin, IEEE, 18(2): 7-12 (2017)
- Sarah Zenasni, Eric Kergosien, Mathieu Roche, Maguelonne Teisseire: Spatial Information Extraction from Short Messages. Expert Syst. Appl. 95: 351-367 (2018)
- Matthijs van Leeuwen: Interactive Data Exploration Using Pattern Mining , In : Interactive Knowledge Discovery and Data Mining : State-of-the-Art and Future Challenges in Biomedical Informatics, LNCS, Springer, pp. 169-182, (2014)

[2] Dans le cadre de travaux menés avec des géographes du projet Herelles sur des problématiques d'aménagement du territoire, un certain nombre de règles peuvent être exprimées (par exemple, *"l'ouverture à l'urbanisation de la zone X de Montpellier est subordonnée à la desserte par des transports collectifs"*). De manière plus précise, deux types de règles peuvent être définies :

- **Règles *objectives* (ou *strictes*)** liées aux SCOTT (schémas de cohérence territoriale) et SRADDET (schéma régional d'aménagement et de développement du territoire).
- **Règles *subjectives* (ou *souples*)** liées aux tendances environnementales, sociétales, politiques, sanitaires, économiques.D

Des travaux menés dans le cadre du projet Herelles et des projets de l'UMR TETIS s'intéressent à l'extraction semi-automatique de ces règles à partir de données textuelles hétérogènes : *données officielles* pour les règles objectives et *données non officielles* (en particulier médias et médias sociaux) pour les règles subjectives.

Contacts et pièces à fournir :

La thèse se déroulera à Caen ou à Montpellier (des déplacements entre les deux sites seront à prévoir). Pour candidater, envoyer les documents suivants (exclusivement au format pdf) à Mathieu Roche (mathieu.roche@cirad.fr) et Bruno Crémilleux (bruno.cremilleux@unicaen.fr) :

- lettre de motivation expliquant vos qualifications, expériences et motivation pour ce sujet ;
- curriculum vitae ;
- relevés de notes de licence 3, de master (ou équivalent pour les écoles d'ingénieur) ;
- lettre de recommandations ou coordonnées de personnes (encadrants de stage, enseignants ou autre personne) pouvant fournir des informations sur vos compétences et votre travail.

Dates importantes :

Date limite de candidature : 30 avril 2021
Notification d'acceptation à l'audition : lundi 3 mai
Auditions le jeudi 6 mai matin
Début de la thèse : octobre 2021

Unités impliquées :

UMR GREYC

Groupe de Recherche en Informatique, Image et Instrumentation
6, boulevard du Maréchal Juin, CS 45053
14050 Caen Cedex 4, France

UMR TETIS

Territoires, environnement, télédétection et information spatiale
500, rue J.F. Breton
34093 Montpellier Cedex 5, France