Stage M2 – S4 2019:

Analyse de données temporelles massives en Science de l'Environnement

Encadrants : Pierre Gançarski (gancarski@unistra.fr) et Agnès Braud (agnes.braud@unistra.fr)

Lieu : ICube – Université de Strasbourg - Illkirch

Contexte de recherche **:**

L'utilisation de techniques de Machine Learning/Data Mining pour l'analyse de séries temporelles est en pleine essor, plus particulièrement en Sciences de l'Environnement. Le projet ANR FRESQUEAU qui s'est déroulé de 2011 à 2015 a permis de construire une base de données hydrologiques importante sur deux grands bassins hydrographiques, correspondant aux districts Rhin-Meuse et Rhône-Méditerranée et Corse et de proposer de nombreuses méthodes d'analyse et de valorisation de celles-ci.

Ces données, couvrant les deux districts, ont été intégrées une base PostgreSQL/PostGIS. Cette base contient 80 tables, dont certaines ont un nombre de lignes important. On trouve notamment plus de cinq cent milliers de lignes correspondant à des mesures climatiques, plus de quatorze millions de mesures pour la physico-chimie, plus de neuf millions d'exploitations dans le registre parcellaire graphique, plus de huit millions de bâtiments et plus d'un million de tronçons hydrographiques. De plus vingt-deux des tables possèdent au moins un attribut représentant une géométrie. Des données physico-chimiques et biologiques couvrant la France entière pour la période 2007-2013, ont également été acquises dans le cadre d'un projet financé par l'ONEMA (2015-2016).

Le projet ADQEAU :

L'objectif du projet ADQEAU, financé par l'ENGEES est de construire des clusters à partir des séquences de données numériques disponibles dans la base de données Fresqueau et d'utiliser ces clusters comme base de construction des classes thématiques par une opération que nous pourrions qualifier de « sémantisation ». Le projet se déroulera en deux phases correspondant chacun à une année universitaire. La première année, il s'agira, grâce à deux stagiaires (dont un thématicien et un informaticien financé directement par l'équipe SDC), de recenser et mettre à jour et en forme les données disponibles. Parallèlement, une adaptation de l'interface MultiCube (permettant le clustering sous contraintes) sera faite. Les premières expériences de clustering sous contraintes seront menées. En deuxième année, il s'agira de valider (grâce un stagiaire thématicien) l'approche de clustering sous contrainte interactif dans le domaine concerné et publier les résultats. Une comparaison pourra être mené avec les résultats obtenus par la recherche de motifs – méthode appliquée précédemment sur les mêmes données. L'encadrement des stagiaires sera assuré par des membres d'ICube et LIVE.

Le stage M2

L'objectif de ce stage est, plus concrètement, de proposer et mettre en place des outils permettant d'interroger les bases de données existantes afin d'extraire des données, qui une fois mises en forme, pourront « alimenter » le logiciel d'analyse de données (FoDoMuST). Il s'agira donc, de créer des modèles (template) de chaînes d'analyse de telles données en python (Un exemple simple : un

template permettra de charger et mettre en forme les données, puis de les transférer à JCL et enfin de mettre en forme pour un affichage).
Pour ce stage, l'étudiant sera aidé par un stagiaire thématicien familier avec les données environnementales. Il le guidera dans la compréhension des problématiques en jeu et leur donnera les conseils « pratiques » à l'implantation de la solution.

Parallèlement, une adaptation des interfaces est nécessaire et sera faite en collaboration avec un étudiant actuellement en stage de M1.

Ce stage sera rémunéré (de l'ordre de 550€ par mois net d'impôt).

Pour postuler ou pour tout renseignement : Pierre Gançarski (gancarski@unistra.fr)