

Sujet de stage de Master 2

Institut de Mathématiques de Bordeaux (IMB) & Laboratoire Bordelais de Recherche en
Informatique (LABRI) - Université de Bordeaux

Analyse de données issues de capteurs du campus de Talence pour l'amélioration des performances énergétiques et des usages des bâtiments.

Encadrants : Jérémie Bigot & Adrien Richou (IMB)

Nicolas Hanusse & Sofian Maabout (LABRI)

`jeremie.bigot@u-bordeaux.fr`; `maabout@u-bordeaux.fr`

Contexte du sujet de stage

La problématique de ce stage se situe autour de la question du traitement temps réel de données de grande taille qui proviennent de bases de données (potentiellement distribuées) qui sont couplées à des systèmes d'objets connectés qui deviennent de plus en plus présents dans notre quotidien (déplacements, activités physiques, météorologie, trafic routier...).

L'application envisagée dans ce stage est l'étude de données issues de capteurs répartis sur les bâtiments du campus de Talence de l'Université de Bordeaux. Ces capteurs permettent de mesurer au cours de temps (à une fréquence élevée) des quantités physiques telles que la température, pression, humidité, niveau de CO2 dans les salles des bâtiments du campus de Talence. Ces capteurs sont connectés à des serveurs qui permettent de stocker ces mesures, en temps réel, dans des bases de données de grande taille.

Un enjeu important (qui fait l'objet de ce stage) est de pouvoir valoriser l'information qui peut être extraite de ces bases de données dans le but d'améliorer les performances énergétiques et les usages des bâtiments du campus de Talence.

Objectifs du stage

Dans une première étape, il conviendra de mettre en place le développement d'outils d'extraction d'information dans ce type de bases de données. Deux approches complémentaires sont envisagées.

- Statistique et mathématiques appliquées. Il conviendra de développer des méthodes d'inférence statistique dans des modèles de données observées dans des réseaux spatio-temporels, et dans un contexte de grande dimension qui tienne compte des bruits d'acquisition et des corrélations entre les observations.
- Fouille de données et visualisation : en plus des statistiques qui répondent à des mesures définies, le domaine de la fouille de données et de la visualisation d'information peut permettre à l'expert de naviguer visuellement entre les nombreuses analyses et représentation des données et d'en faire une synthèse.

Dans une deuxième étape, il est envisagé de considérer la problématique du traitement et stockage d'un flux massif de données. Les technologies liées à l'informatique en nuage ("Cloud Computing") ont permis le développement de nouvelles techniques de programmation génériques (Hadoop, Spark) permettant de traiter des données distribuées avec des moyens de calcul distribué. Les axes de travail

envisagés portent sur :

- la définition d'une algorithmique à même de garantir une précision sur les résultats avec peu de mémoire (modèle flux) ;
- le passage des algorithmes de fouille de données ou d'apprentissage dans le cadre du calcul distribué ;
- la définition de nouvelles représentations ou structures de données (réseaux et données spatio-temporelles) s'adaptant à une algorithmique distribuée ;
- proposer des paradigmes de programmation simples et efficaces. En effet, les solutions actuelles de calcul distribué (Map Reduce, Spark) sont très lentes par rapport au calcul sur des machines parallèles puissantes.

Compétences attendues du/de la candidat(e)

- Master 2 ou 3ème année d'école d'ingénieur en mathématiques appliquées ou informatique.
- Spécialisation en modélisation statistique et/ou algorithmique distribuée et informatique pour l'analyse de données massives.
- Connaissance de logiciels pour l'analyse de données.

Le début du stage est possible **dès février 2018 pour une durée de 5 à 6 mois**, et il est rémunéré selon la gratification correspondant à l'indemnité de stage en vigueur en 2018.

En fonction de son déroulement, le **stage pourra déboucher sur une thèse** autour de la thématique de

“L'extraction de connaissances dans les données massives et hétérogènes issues d'un réseau de capteurs”

dans le cadre de l'axe de recherche “Données massives et hétérogènes” du cluster SysNum (initiative d'excellence de l'Université de Bordeaux) :

<http://idex.u-bordeaux.fr/fr/n/Defis-scientifiques/Systemes-numeriques/r3013.html>