

Proposition de thèse

Titre	Modélisation et Intégration des Données Multi-vues : détection de données aberrantes et extraction de règles.	
(Co)-Directeurs	Stéphane Lecoecueche	E-mail : stephane.lecoecueche@imt-lille-douai.fr
	Arnaud Doniec	E-mail : arnaud.doniec@imt-lille-douai.fr
(Co)-Encadrants	Jerry Lonlac	E-mail : jerry.lonlac@imt-lille-douai.fr
	Alexis Lesage	E-mail : a.lesage@intent-technologies.eu
Laboratoire	IMT Lille Douai, CERI SN	Web : https://imt-lille-douai.fr/recherche-et-innovation/systemes-numeriques/
Equipe	McLeod	Web : https://imt-lille-douai.fr/recherche-et-innovation/systemes-numeriques/mcleod/
	Intent	Web : https://intent.tech/
Financement prévu	ANR AI@IMT	
	Région – Autre <input type="checkbox"/>	Contrat de recherche <input checked="" type="checkbox"/> Préciser : Intent
Financement acquis ? <input type="checkbox"/>	Contrats de Recherche <input type="checkbox"/> Préciser	Autre <input type="checkbox"/> Préciser

Résumé du sujet :

La société Intent développe une plateforme de services et d'agrégation de données, dédiés aux bâtiments et à la ville. Cette offre constitue une API unique et complète de 120 services déjà connectés autour d'une variété importante d'enjeux verticaux : énergie, autonomie, maintenance. La plateforme d'Intent collecte et centralise ainsi toutes les données d'un bâtiment et facilite la gestion des services apportés aux occupants, exploitants et gestionnaires. La plupart du temps, ces différents acteurs partagent des visions voire des informations différentes du bâtiment : un gestionnaire de patrimoine foncier n'aura pas forcément la même connaissance qu'un prestataire de maintenance : la nomenclature utilisée, le contenu des bases, leur niveau de détails peuvent par exemple varier. Ainsi, pour une même réalité sur le terrain, plusieurs sources (capteurs, déclarations d'occupants, saisie manuelle d'opérateurs de maintenance) et vues distinctes peuvent exister au sein de la plateforme d'Intent. Ainsi, Intent est confronté à la mise en place de moteurs d'appariement, actuellement réalisée grâce à une expertise humaine. Cependant, face à l'accroissement de l'activité, du volume de données hébergées et de la multiplication des vues d'utilisateurs, il est indispensable pour l'entreprise de permettre ces appariements de façon plus automatique.

Dans ce contexte, l'objectif de cette thèse est de contribuer au développement de nouveaux outils de fouille de données capables de composer avec plusieurs vues, tout en assurant la fiabilité des données et l'appariement d'information. Deux verrous seront abordés dans cette thèse :

- la collecte de données hétérogènes et la détection de données aberrantes,
- l'extraction de règles à partir d'un apprentissage multi-vue.

Les objectifs du premier verrou concernant (1) la détection de changements dans les caractéristiques des données et de leur dépendance (2) la caractérisation des motifs et des anomalies (3) l'annotation des données



correspondantes. Ces objectifs requièrent le développement de nouvelles méthodes d'évaluation de la qualité informative et de la détection de changement de distributions. Ces travaux pourront s'appuyer sur les savoir-faire du laboratoire autour des techniques d'apprentissage pour la détection de nouveautés [1, 2], mais aussi sur les outils de la modélisation de comportements dynamiques [3] qui peuvent être adaptés pour traiter des données fonctionnelles ou des données graduelles [4].

L'objectif visé du second verrou est (1) d'explorer la construction, à partir de motifs isolés, de règles d'association dans le contexte de données hétérogènes multi-vues et (2) d'étudier leur apport pour l'appariement d'information décrivant une même réalité sur le terrain. Pour atteindre les objectifs susmentionnés, nous nous appuyerons sur nos travaux autour de l'extraction des motifs intéressants dans des masses de données [4, 5] et autour du clustering des données [1, 2]. Les travaux menés dans les deux axes visent ainsi à mettre en place de nouvelles méthodes permettant de fiabiliser les échanges de données au sein de la plate-forme Intent.

Le travail de thèse est structuré autour des points suivants :

- Découvrir la plateforme de données Intent et s'appropriier la problématique (T0+3mois d'immersion dans l'entreprise),
- Contribuer à un état de l'art sur la fouille de données hétérogènes dans un contexte multi-source (T3+6mois),
- Concevoir et implémenter des algorithmes de marquage de données à partir d'une mise en correspondance, d'une détection de changement et du suivi des critères d'information (T3+18mois),
- Contribuer à un état de l'art sur l'extraction des règles d'association à partir des données multi-vues (T12+6 mois),
- Construire une méthodologie d'extraction de règles d'association dans les bases de données multi-vues (T12+12 mois),
- Implémenter et tester les méthodes d'agrégation des données multi-sources sur la plateforme Intent (T18+12 mois),
- Finaliser un livrable logiciel et un manuscrit (TF).

Bibliographie :

- [1] H. A. Boubacar, S. Lecoeuche, S. Maouche, (2008). SAKM: Self-adaptive kernel machine A kernel-based algorithm for online clustering. *Neural Networks*, 21(9), 1287-1301
- [2] E. Ghorbel, J. Boonaert, R. Bouteau, S. Lecoeuche, X. Savatier, (2018). An extension of kernel learning methods using a modified log-euclidean distance for fast and accurate skeleton-based human action recognition. *Computer Vision and Image Understanding*, 175, 32-43.
- [3] L. Bako, K. Boukharouba, E. Duviella, S. Lecoeuche, (2011). A recursive identification algorithm for switched linear/affine models. *Nonlinear Analysis: Hybrid Systems*, 5(2), 242-253
- [4] J. Lonlac, A. Doniec, M. Lujak, S. Lecoeuche, (2020, September). Mining Frequent Seasonal Gradual Patterns. In *International Conference on Big Data Analytics and Knowledge Discovery* (pp. 197-207). Springer, Cham
- [5] J. Lonlac, E. Mephu Nguifo. A novel algorithm for searching Frequent Gradual Patterns from an Ordered Data Set. *Intelligent Data Analysis (IDA)*, 24(5), 1029–1042, 2020.

Profil du candidat souhaité :

Le candidat recherché pour ce sujet devra avoir un Master 2 ou un diplôme d'ingénieur et avoir une forte expertise en informatique, en particulier en intelligence artificielle (Data mining, Machine Learning), en modélisation des données hétérogènes issues de différentes sources, mais aussi en bases de données. Des



IMT Lille Douai
École Mines-Télécom
IMT-Université de Lille

connaissances dans le domaine du bâtiment seront appréciées afin de pouvoir mieux appréhender les données et les différents services hébergés par la société Intent.

Candidature :

Les candidats sont invités à envoyer CV, lettre de motivation et résultats de master à sandra.andrzejewski@imt-lille-douai.fr, avec copie aux directeurs et encadrants.