AVIS DE PRESENTATION DE THESE EN SOUTENANCE POUR L'OBTENTION DU DIPLOME NATIONAL DE DOCTEUR

Monsieur Salah Eddine BOUKHETTA

Présentera ses travaux intitulés :

« Analyse de séquences avec GALACTIC – Approche générique combinant analyse formelle des concepts et fouille de motifs »

Spécialité : Informatique et Applications

Le 30 aout 2022 à 15h00

Lieu :

La Rochelle Université
Pôle Communication, Multimédia et Réseaux
Amphithéâtre Michel Crépeau
44 Av. Albert Einstein
17000 LA ROCHELLE

Composition du jury :

Mme BERTET Karell	**Maîtresse de conférences, HDR, La Rochelle Université**
M. DEMKO Christophe	**Maître de conférences, La Rochelle Université**
M. FAUCHER Cyril *(Invité)*	**Maître de conférences, La Rochelle Université**
M. FERRÉ Sébastien	**Professeur, Université Rennes 1**
Mme GIRARD Nathalie	**Maîtresse de conférences, Université Rennes 1**
M. GUYET Thomas	**Chargé de recherche, INRIA de Lyon**
M. HUCHARD Marianne	**Professeure, Université Montpellier 2**
M. KAYTOUE Mehdi *(Invité)*	**HRD, Entreprise Infologic**
M. LEJEUNE Gaël	**Maître de conférences, Sorbonne Université**
Mme LE BER Florence	**Professeure, ENGEES**

Résumé :

Les données de type séquences sont à nos jours utilisées dans beaucoup de domaines dans le but de mieux les analyser et d'en extraire des connaissances. Une séquence est une suite d'éléments ordonnés comme les trajectoires de déplacement ou les séquences d'achats de produits dans un supermarché. Il existe plusieurs types des séquences trois types de séquences, les séquences simples, les séquences temporelles et les séquences d'intervalles. La fouille de séquences est un domaine de la fouille de données qui vise à extraire des motifs séquentiels fréquents à partir d'un ensemble de séquences, où ces motifs sont le plus souvent des sous-séquences communes. Un motif séquentiel peut être une sous-séquence, et il est fréquent s'il est partagé par un nombre suffisamment représentatif de données. Le support est une mesure monotone qui définit la proportion de données partageant un motif séquentiel. Plusieurs algorithmes ont été proposés pour l'extraction des motifs séquentiels fréquents. Avec l'évolution des capacités de calcul, la tâche d'extraction des motifs séquentiels fréquents est devenue plus rapide. La difficulté réside alors dans le trop grand nombre de motifs séquentiels extraits, qui en rend difficile la lisibilité et donc l'interprétation. On parle de déluge de motifs. Une première approche pour résoudre ce problème consiste à réduire les motifs fréquents générés aux seuls motifs fermés qui portent la même information. Il a été observé que l'ensemble de tous les motifs fermés peut être organisé dans une structure appelée treillis. Cette structure est la base de l'Analyse Formelle de Concepts (AFC) qui est un domaine d'analyse de données permettant d'identifier des relations dans l'ensemble de données. L'AFC est classiquement conçue pour traiter des données décrites par des ensembles d'attributs, donc des données binaires. Le formalisme des structures de motifs et la navigation conceptuelle abstraite étendent l'AFC pour traiter des données complexes comme les séquences. Inspiré des structures de motifs, l'algorithme NextPriorityConcept propose une approche d'extraction de motifs pour des données hétérogènes et complexes. Cet algorithme permet un calcul de motifs génériques à travers des descriptions spécifiques d'objets par des prédicats monadiques. Il propose également de raffiner un ensemble d'objets en un ensemble plus petit à travers des stratégies d'explorations spécifiques de l'utilisateur, ce qui permet de réduire le nombre de motifs et ainsi limiter le déluge de motifs générés. La plateforme GALACTIC implémente l'algorithme NextPriorityConcept et propose un écosystème d'extensions pour le traitement de données. Dans ce travail, nous nous intéressons à l'analyse de données séquentielles en utilisant GALACTIC.

Nous proposons plusieurs descriptions et stratégies adaptées aux séquences simples, temporelles et d'intervalles. Les descriptions sont basées sur les sous-séquences communes maximales avec des descriptions plus spécifiques aux types des séquences. Nous proposons une stratégie naïve permettant la génération de tous les concepts et des stratégies plus élaborées permettant de réduire la taille du treillis. Nous proposons des mesures de qualité non supervisées pour pouvoir comparer entre les treillis obtenus. Une analyse qualitative et quantitative est mené sur des jeux de données réels et synthétiques afin de montrer l'efficacité de notre approche.