Projet de thèse doctorat : Fouille de séquences sémantiques

Directeur de thèse : Thomas Devogele, professeur des universités

Encadrants : Verónica Peralta, maître de conférences, Cyril de Runz maîtres de conférences HDR

Localisation du doctorant : Université de Tours, laboratoire d'informatique : LIFAT, antenne de Blois

Les séquences ou trajectoires sémantiques sont des suites de valeurs sémantiques ayant une durée et chronologiquement ordonnées. L'ensemble de ces valeurs peut être décrit à l'aide d'une ontologie. Ces séquences représentent des processus divers : trajectoires de vie, déplacements journaliers, dossiers patients et flux d'activités diverses (par ex. étapes dans la production de produits, exercices d'e-learning, requêtes dans un système d'information, chansons d'une playlist...).

L'analyse de séquences sémantiques (Parent et al. 2013) permet de répondre à diverses problématiques sociétales, industrielles ou individuelles, par exemple : la détection de comportements dangereux (ex. déplacements à risque, notamment d'enfants, suites dangereuses de prises de médicaments), la détection de difficultés et goulots d'étranglement (notamment quand il s'agit d'activités répétitives), l'apprentissage de patrons de comportement (pour ensuite créer des groupes, observer des caractéristiques communes, recommander des actions, apprendre des intérêts), et la comparaison de comportement experts et novices (pour qualifier, recommander et proposer des plans d'amélioration).

En 2021, Clément Moreau (Moreau 2021) a soutenu sa thèse sur la fouille de séquences de mobilité sémantique. Il a proposé des mesures pour la comparaison, l'analyse et la découverte de comportements d'humains. Ces travaux de recherche ont ouvert de nombreuses perspectives très prometteuses.

L'objectif de cette thèse est d'étendre ces travaux et de répondre à une partie des verrous soulevés par ces perspectives :

- **Généricité** : La thèse de Clément Moreau s'intéressait principalement aux séquences sémantiques représentant des déplacements humaines. Une question intéressante concerne *comment étendre ces travaux à d'autres types de données*, notamment pour prendre en compte des activités complexes et en présence d'incomplétudes. Un couplage avec la dimension géographique est possible.
- **Mesure de similarité** : Le calcul de similarité entre activités et entre séquence est un point clé pour l'analyse des séquences. Clément Moreau a proposé deux mesures, CED et FTH basées respectivement sur la distance d'édition (Wagner et Fischer 1974) et la distance de Hamming (Hamming 1950). Une étude comparative approfondie est encore à réaliser. Plus particulièrement, *quelles caractéristiques sont à privilégier pour choisir et régler les paramètres de la mesure de similarité ?* Comment mesurer la similarité entre des séquences incomplètes ou de durées différentes ?
- **Langage d'interrogation** : Ces mesures doivent pouvoir être reprises dans le cadre d'un langage d'interrogation permettant d'extraire des séquences proches d'un patron générique.
- **Analyse visuelle interactive** : Lors de la thèse de Clément Moreau, un outil de visualisation : SIMBA a été proposé. Cet outil est complexe, il doit être amélioré afin de proposer des résumés (patterns) plus simples des séquences d'un même cluster et ainsi favoriser

l'explicabilité du processus. De même, l'analyse doit être plus interactive et mieux intégrer les préférences utilisateurs.

Financement : CDD de 3 ans (début octobre 2023) [contrat doctoral de droit public](#) avec possibilité d'enseignement.

Pour candidater, veuillez envoyer un CV et une lettre de motivation avant le 26 mai 2023 à :

- Thomas Devogele : thomas.devogele@univ-tours.fr
- Verónica Peralta : Veronika.Peralta@univ-tours.fr
- Cyril de Runz : cyril.derunz@univ-tours.fr

Références

Hamming, R. W (1950) Error detecting and error correcting codes. The Bell system technical journal, 29(2):147–160.

Moreau, C. (2021). Fouille de séquences de mobilité sémantique : sur l'élaboration de mesures pour la comparaison, l'analyse et la découverte de comportements (Thèse en informatique de l'Université de Tours). [PDF](#)

Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., ... & Yan, Z. (2013). Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, 45(4), 1-32.

Wagner, R. A. et Fischer M. J. (1974) The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173.