Similarité de séquences sémantiques

Hiba Merakchi

Similarité de séquences sémantiques

Hiba MERAKCHI

Laboratoire LIFAT, Université de Tours
Campus Universitaire de Blois, 3 place Jean Jaurès
41000 Blois, France

hiba.merakchi@univ-tours.fr

MOTS-CLÉS : *fouille de données, séquences sémantiques, trajectoires sémantiques, distance d'édition, logique floue, ontologie*

RÉSUMÉ: *Les séquences sémantiques offrent un potentiel pour comprendre et anticiper les comportements humains. Cette recherche explore deux nouvelles mesures de similarité, la Distance d'Édition Contextuelle (CED) et la Distance de Hamming Temporelle Floue (FTH). Nos travaux visent à améliorer ces méthodes, développer un langage de requêtes pour trouver des séquences similaires, améliorer l'explicabilité des résultats, et valider leur applicabilité dans divers contextes de données.*

ENCADREMENT. *Thomas DEVOGELE, Veronika PERALTA, Cyril DE RUNZ.*

1. Introduction

Les séquences sémantiques désignent des ensembles d'éléments chronologiquement ordonnés, chaque élément ayant une sémantique et potentiellement une durée. Ces éléments, définis et interprétés à l'aide d'une ontologie, peuvent représenter une multitude d'événements, actions et activités humaines. Les séquences sémantiques permettent donc de représenter divers processus et enchaînements d'activités humaines, par exemple, des trajectoires de mobilité, des parcours de vie, des dossiers patients et des flux d'activités diverses (comme des étapes dans les chaînes de productions, des exercices d'e-learning, des requêtes dans un système d'information, ou des chansons d'une playlist).

Par exemple, la séquence de la figure 1 représente la mobilité d'une personne pendant une partie de la journée. La personne est chez elle de 4h à 8h du matin. Elle

prend le bus de 8h à 8h30 pour se rendre au travail. À 16h30, elle finit le travail et marche pendant 15 minutes pour aller à la piscine, où elle nage jusqu'à 18h. Ensuite, elle marche pendant 30 minutes et fait du télétravail de 18h30 à 19h30.

L'analyse de ces séquences sémantiques, comme discutée par (Parent C., 2013), ouvre la voie à une multitude d'applications aussi variées que cruciales pour la société contemporaine. En effet, au-delà de leur simple description, ces séquences offrent un potentiel considérable pour répondre à des défis sociétaux, industriels et individuels. L'analyse des séquences sémantiques offre également la possibilité d'apprendre des modèles de comportement. Cette capacité d'apprentissage permet de créer des groupes homogènes, d'observer des caractéristiques communes et même de recommander des actions ou d'anticiper des intérêts potentiels.

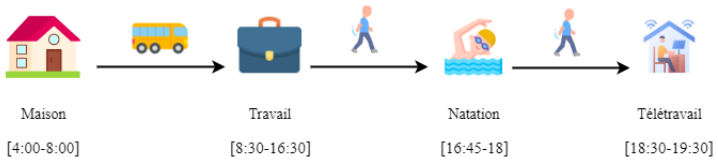

Figure 1: Exemple de séquence avec annotations sémantiques (Moreau C., 2021).

2. Motivations : Mesures de similarité entre séquences

La thèse de (Moreau C., 2021), portant sur l'exploration des séquences de mobilité sémantique, introduit deux nouvelles mesures de similarité de séquences : la Distance d'Édition Contextuelle (CED) et la Distance de Hamming Temporelle Floue (FTH). Ces mesures sont inspirées des méthodes existantes. Elles sont particulièrement adaptées à l'analyse des séquences grâce à l'utilisation d'ontologies et de logique floue, et sont au coeur d'un processus de clustering et d'une méthodologie d'analyse de séquences. En mettant en œuvre cette méthodologie sur des jeux de données synthétiques et réels issus de différents domaines de la mobilité, Clément Moreau a pu améliorer significativement la capacité à interpréter et découvrir des comportements.

CED étend la Distance d'Édition en tenant compte du contexte, qui désigne le contenu sémantique ou une partie de la séquence. Elle repose sur le produit de deux fonctions. La première mesure la similarité sémantique à l'aide d'une ontologie. La deuxième prend en compte l'écart de position.

Ainsi, elle répond à des propriétés mathématiques comme l'homogénéité sémantique, la temporalité des activités, le décalage temporel, la permutation d'activités et la redondance d'activités (Moreau C., 2021).

FTH, quant à elle, est une extension floue de la distance de Hamming, pour les séquences sémantiques-temporelles. Cette mesure améliore la capacité de comparer des séquences de durées égales en introduisant une fenêtre temporelle floue pour gérer

les distorsions temporelles telles que les décalages et les permutations. FTH garantit également d'autres propriétés telles que la temporalité et l'homogénéité sémantique.

3. Actions réalisées : Uniformisation et analyse de sensibilité

Dans le cadre de notre recherche en cours, visant à améliorer les méthodes CED et FTH, nous avons commencé par une étude comparative, en tenant compte de plusieurs axes : la dimension sémantique, contextuelle, floue et temporelle. Nous souhaitons déterminer quelles caractéristiques sont à privilégier lors du choix et du réglage des paramètres de la mesure de similarité.

En premier temps, nous avons travaillé sur l'unification des mesures FTH et CED en intégrant la logique floue dans CED. Plus concrètement, nous utilisons une fonction floue trapézoïdale comme fonction de contexte (plutôt que la fonction exponentielle utilisée par CED). Nous avons également commencé à tester l'impact du choix du support de la fonction floue qui caractérise et contrôle le degré de proximité temporelle entre les activités sur les deux mesures et ultimement sur la comparaison des séquences.

De plus, nous avons testé les performances de CED et FTH sur des séquences avec des activités de même durée et avons constaté que les deux méthodes donnent des résultats similaires, comme attendu. Cependant, FTH pourrait être plus intéressante dans le cas des séquences comportant des activités de durées différentes. Cela reste à explorer.

4. Actions futures : Recherche de sous-séquences, explicabilité et généricité

En nous inspirant des travaux sur les requêtes floues par l'exemple (Moreau A., 2018, Smits G., 2013), nous souhaitons définir un langage de requêtes favorisant la recherche de ***top-k*** séquences similaires à un motif (une sous-séquence). Par exemple, si nous cherchons à trouver toutes les personnes ayant travaillé environ 8 heures et qui ont marché pendant à peu près une heure pour s'y rendre et en revenir (figure 2). Nous désirons retourner des séquences où les horaires peuvent varier, les activités peuvent être segmentées, voire inversées ou remplacées par des activités proches. L'utilisation des distances classiques comme la distance d'édition ne permettrait pas de trouver la séquence de la figure 3. En effet, la marche à pied est remplacée par la trottinette et l'activité "Travail" a été segmentée par l'activité "Repas".

Figure 2: Exemple de sous-séquence recherchée.

La séquence illustrée en figure 3 pourrait être retournée par une telle requête. Elle comprend des déplacements doux (en trottinette, à pied) ainsi que des périodes de travail pas nécessairement de la même durée recherchée ni dans le même ordre.

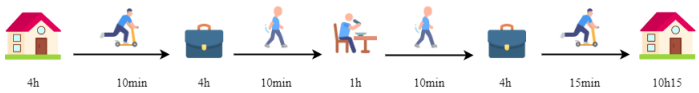

Figure 3: Exemple de séquences retournées.

L'objectif est donc de concevoir un langage permettant de définir des critères de similarité pour identifier des sous-séquences proches en termes de contexte temporel et sémantique, tout en préservant des propriétés de permutation, répétition et homogénéité sémantique. Cela faciliterait l'analyse et la compréhension des habitudes de mobilité d'une personne.

De plus, nous allons nous atteler à développer davantage la partie ***explicabilité*** de nos méthodes, en cherchant à identifier des motifs fréquents ou centraux représentatifs de sous-ensembles (par exemple, des *clusters*) de séquences. Ces avancées nous permettront d'offrir à l'utilisateur une compréhension plus approfondie des résultats et de faciliter leur interprétation dans divers contextes d'application.

Nous envisageons également d'étendre l'utilisation de CED et FTH à différentes sources de données pour valider leur généricité. Nous utiliserons des données d'activités provenant de maisons intelligentes, des Enquêtes Ménages Déplacements (EMD), des listes de lecture de chansons, ainsi que des fichiers de logs pour tester, confirmer et valider l'applicabilité de ces méthodes dans divers contextes.

Bibliographie

Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., Macedo, J., Pelekis, N., et al. (2013). Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, *45*(4), 1–32. ACM New York, NY, USA.

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 de doctorat, Université de Tours.

Moreau, A., Pivert, O., Smits, G. (2018). Fuzzy query by example. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (pp. 688–695).

Smits, G., Pivert, O., Girault, T. (2013). Reqflex: fuzzy queries for everyone. *Proceedings of the VLDB Endowment*, *6*(12), 1206–1209. VLDB Endowment.