Extraction de motifs de trajectoires sémantiques similaires

Clément Moreau, Thomas Devogele, Laurent Etienne

Extraction de motifs de trajectoires sémantiques similaires

Clément Moreau[1], **Thomas Devogele**[1], **Laurent Etienne**[1]

Laboratoire d'Informatique Fondamentale et Appliquée de Tours
64, avenue Jean Portalis, 37200 Tours France
{clement.moreau,thomas.devogele,laurent.etienne}@univ-tours.fr

RÉSUMÉ. *La compréhension fine des déplacements des individus nécessite une modélisation sémantique riche de leurs activités. Or, il est maintenant possible d'extraire les mobilités et les activités des individus à l'aide d'informations contextuelles ou de capteurs. Une fois enrichies sémantiquement, ces mobilités peuvent être comparées selon une mesure de proximité spatiale, temporelle et sémantique, puis regroupées en clusters de trajectoires similaires. Afin de résumer ces déplacements similaires, des motifs synthétisant ces clusters peuvent être induits. Cet article présente une méthodologie pour extraire ces motifs à partir de trajectoires sémantiquement riches ; fort utile aux experts pour mieux analyser les déplacements. Dans cet objectif, cet article propose de mettre en lumière, d'étendre et de relier un grand nombre d'outils de la fouille des trajectoires. Cette méthode est générique et s'applique à nombre de domaines d'études tels que le tourisme, la sociologie, l'épidémiologie ou l'urbanisme.*

ABSTRACT. *Thanks to the growth of Internet of things and use of various sensors embedded in smartphones, individuals can be tracked and monitored all day long. However, the fine understanding of their activities requires semantic models and contextual information. The users' trajectories can then be enhanced with semantic meaning. Trajectories can be compared using their spatial, temporal or semantic components. Similar users' behaviour can then be clustered to derive movement patterns. This article presents a new methodology to extract movement patterns from semantic trajectories. These patterns are helpful for experts working on movement analysis in various fields such as tourism, sociology, epidemiology or urban planning.*

MOTS-CLÉS : *Trajectoires sémantiques, Clusters de trajectoires, Pattern de déplacement, Mesure de similarité, Fouille de trajectoires*

KEYWORDS: *Semantic trajectory, Trajectory clustering, Moving Object patterns, Similarity measure, Trajectory datamining*

1. Introduction et contexte

La compréhension de la mobilité humaine est un enjeu dans de nombreux domaines comme l'urbanisme, le tourisme, la sociologie, ou encore l'analyse de propagation des virus. Un résultat primordial donné par (González *et al.*, 2008) présente que malgré la complexité et diversité des trajectoires, la mobilité humaine présente un degré élevé de régularité temporelle et spatiale ce qui indique que les déplacements individuels suivent des modèles reproductibles simples. Cependant, même si une clarté semble se dessiner autour du caractère prédictible des trajectoires humaines et que les résultats précédemment énoncés fournissent une réponse rassurante quant à la compréhension des schémas qui animent les déplacements, ceux-ci demeurent éloignés des interrogations véritables que porte la mobilité humaine qui sont le sens et la contextualisation (Renso, Trasarti, 2013 ; Yan, 2009 ; Parent *et al.*, 2013).
C'est au sein de cette brèche sémantique que nous posons ici notre propos. Au cours de notre exposé nous illustrerons notre sujet à l'aide d'exemples tirés de deux projets applicatifs : SmartLoire et Mobi'Kids, respectivement une plateforme de recommandation pour le tourisme sur mesure en région Centre - Val de Loire, et une étude sociologique visant à comprendre les conditions de mobilités quotidiennes des enfants dans un contexte impulsé par les enjeux de la ville durable et des modes alternatifs de déplacements. L'objectif consiste à vérifier l'hypothèse qu'il existe des formes de "cultures éducatives urbaines" variant selon les lieux de vie, les situations sociales et les modes de vie.
L'abondance de ressources complémentaires au GPS tend à affiner la connaissance que nous avons de la mobilité des individus. Dès lors, il convient d'adopter une modélisation des trajectoires en adéquation avec cette richesse sémantique disponible. De cette modélisation découle alors une métrique permettant de comparer deux trajectoires en vue d'établir un algorithme de partitionnement de données (ou clustering) afin de regrouper les motifs (ou pattern) de comportements similaires. Une dernière étape se traduit par une représentation synthétique des motifs extraits.
L'existence d'une telle chaîne de traitement, analogue à la Figure *1.* répond entre autres à trois des challenges majeurs proposés par (Ferrero *et al.*, 2016), soient : Comment représenter les informations contextuelles et hétérogènes au sein des trajectoires sémantiques (*Multiple Aspect Representation*) ? Comment enrôler les dimensions spatiale, temporelle et sémantique au sein d'une même métrique et établir des partitions de trajectoires proches (*Similarity Analysis and Data Mining*) ? Et enfin est-il possible de résumer une partition de trajectoires proches en un motif synthétique (*Vizualisation*) ? Cet article offre un éclairage et des premières réponses à ces problèmes complexes en mettant en lumière un ensemble de méthodes afin de décrire le processus proposé et en adoptant des solutions issues de la littérature. Elles sont illustrées par des cas d'école.
L'article est organisé comme suit : Dans la section 2 , les modèles de trajectoires sémantiques sont présentés. En section 3, nous passons en revue les différentes

mesures de proximité puis présentons une distance permettant de comparer deux trajectoires au sein de notre modèle sémantique, cette distance est reprise en section 4 où sont abordées les problématiques de fouille telles que le partitionnement des trajectoires et l'extraction de motifs fréquents.

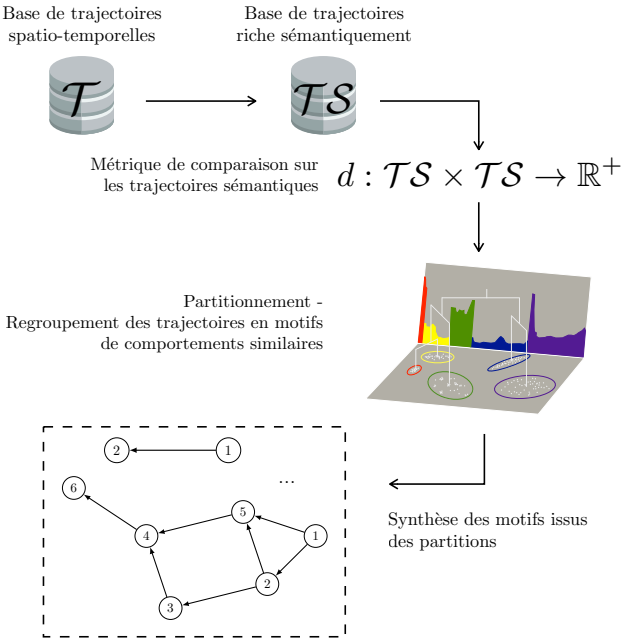

FIGURE *1. Chaîne de traitement pour l'extraction de motifs de trajectoires sémantiques similaires*

2. Modélisation des trajectoires sémantiques

2.1. Les différentes représentations sémantiques des trajectoires

La notion de sémantique au sein des trajectoires spatio-temporelles naquit de la volonté d'introduire une dimension contextuelle au déplacement observé afin de cerner au mieux sa nature. Elle émergea également dans une perspective d'intelligibilité des données afin de réduire le temps de calcul lors des opérations de fouille et d'analyse de données et de rendre l'interrogation plus naturelle. C'est dans cette seconde optique que (Alvares *et al.*, 2007) présente un modèle permettant d'extraire les points de séjour et déplacements (stops and moves) d'une trajectoire. Les lieux d'arrêt et points de séjour sont ensuite enrichis sémantiquement par un label (hôtel, musée, etc...) en vue d'être analysés de façon

plus haut niveau et non plus seulement géométrique. (Spaccapietra *et al.*, 2008) se réapproprie les notions de Stop and Move en vue d'établir un formalisme de conception de référence pour la modélisation sémantique des trajectoires.
Certains auteurs ont tenté d'élargir la conception de Spaccapietra et al., dans (Yan *et al.*, 2010) la trajectoire sémantique est représentée comme une séquence d'épisodes sémantiques. De nombreux modèles comme (Noël *et al.*, 2015 ; Beber *et al.*, 2017) reprennent cette modélisation en adoptant des codes du modèle CONSTAnT établi dans (Bogorny *et al.*, 2014) qui possède un pouvoir d'expression élevé. Bien souvent la fouille de données exige la prise en compte de paramètres environnementaux (météo, aménagement urbain, lieux d'intérêt) mais aussi thématiques (Base d'horaires des transports, événements, activités, routines et comportements). C'est notamment en réponse à ces besoins réels que ces modèles plus élaborés, comme CONSTAnT, ont vu le jour. Ils incluent plusieurs notions formalisées telles que les : lieux géographiques et sémantiques, événements, objectifs, activités, environnement et comportement dans l'espoir de devenir une référence dans la conception des trajectoires sémantiques.
Cependant, même si la représentation de Bogorny et al. est particulièrement féconde, elle échoue dans la réponse à l'ensemble des conditions posées par (Yan, 2009) : L'alimentation du contexte des trajectoires par des ontologies à des fins d'inférences est manquante, on ne considère aucune hiérarchie de concepts permettant une généralisation ou d'établir une distance entre deux entités.
Des vétilles énoncées sur CONSTAnT, nous retiendrons le défaut majeur suivant : malgré la volonté de vouloir s'insérer dans une démarche forte pour la fouille de données, le modèle ne fournit malheureusement pas de métrique de comparaison nécessaire. Hormis ce point, nous soutenons la richesse d'expressivité sémantique offerte par le modèle de Borgony et al. et pensons que des modifications telles que celles suggérées dans (Parent *et al.*, 2013 ; Gibert *et al.*, 2013) comme la prise en compte d'ontologies pour l'analyse de données et la généralisation de concepts sont des points-clés à son amélioration.

2.2. Modèle de trajectoire riche sémantiquement

Le modèle des trajectoires riches sémantiquement retenu s'appuie sur la notion d'activité. Il reprend les concepts du modèle CONSTAnT et lui associe des informations ontologiques. La figure 2. synthétise ce modèle. Les trajectoires sémantiques des individus sont des suites ordonnées temporellement d'activités sans recouvrement temporel. Une activité est décrite selon les trois dimensions : sémantique, spatiale et temporelle. Les concepts ontologiques sont organisés hiérarchiquement selon un arbre ou un treillis ce qui est indispensable pour définir une notion de proximité entre concepts. Une activité est rattachée à un concept ontologique, par exemple les concepts "Natation", "Basket-ball" et "Course" sont des sous-concepts du concept "Activité sportive". Le concept `NULL` est autorisé, il représente une activité inconnue ou une activité d'attente. Les activités sont associées à des trajectoires. Ces objets spatio-temporels re-

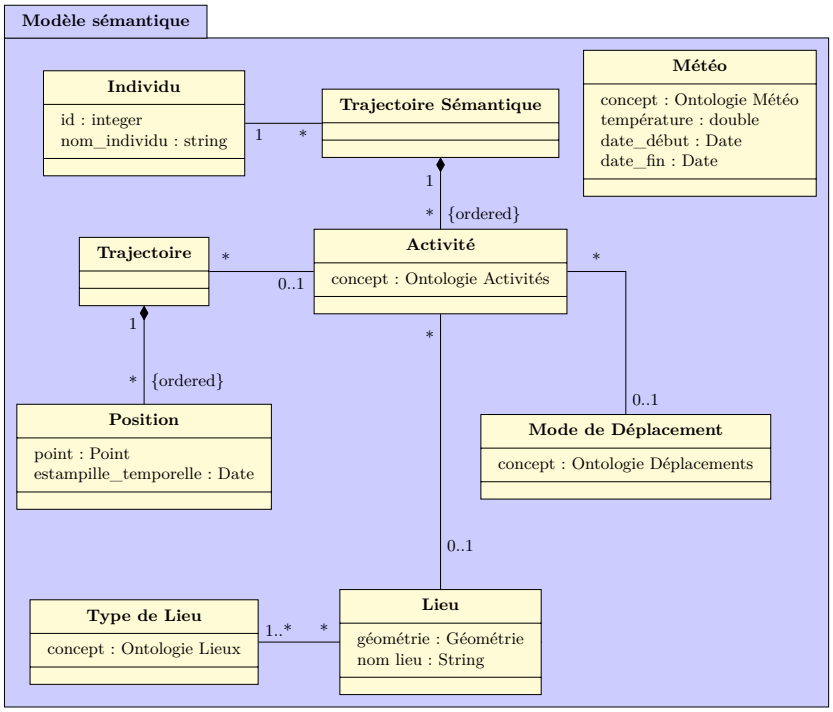

FIGURE *2. Modélisation UML des trajectoires sémantique des individus*

groupent des positions ordonnées temporellement sous la forme de suites de triplets (x, y, t). Une activité est statique ou mobile. Les trajectoires d'activités statiques regroupent uniquement un couple de positions avec des coordonnées géographiques identiques et deux estampilles temporelles différentes. Les trajectoires d'activités mobiles relient à l'inverse des suites de positions avec des coordonnées différentes. La dimension temporelle des activités est donc portée par leur trajectoire. Le début de la période est l'estampille temporelle de la première position. De même, la fin de la période est l'estampille de la dernière position. Comme pour le modèle SeMiTri de (Yan *et al.*, 2011), l'alternance entre les activités statiques et mobiles n'est pas obligatoire. Un mode de déplacement ("à pied", "à vélo", "en voiture", etc.) est défini également à l'aide d'un concept ontologique et est réservé aux seules activités mobiles.
Les activités sont potentiellement attachées à des lieux et sont décrites par un nom (P_1 et G_1 par exemple), une géométrie et sont associées aux concepts d'une ontologie de lieux par des liens `is_a`. Ainsi, P_1 est une "Piscine" et G_1 est un "Gymnase". Ces deux concepts sont reliés dans une hiérarchie de concepts au concept "Équipement Sportif". Il faut aussi noter qu'une activité sportive

ne se déroule par obligatoirement dans un lieu de type "Équipement sportif". En fonction des applications, la météorologie joue un rôle important. Dans ce cas, la météo est relatée par une suite de périodes avec deux valeurs : une température et un concept ontologique temps.
Les 7 trajectoires schématiques de la figure 3. vont servir d'exemple fil rouge à

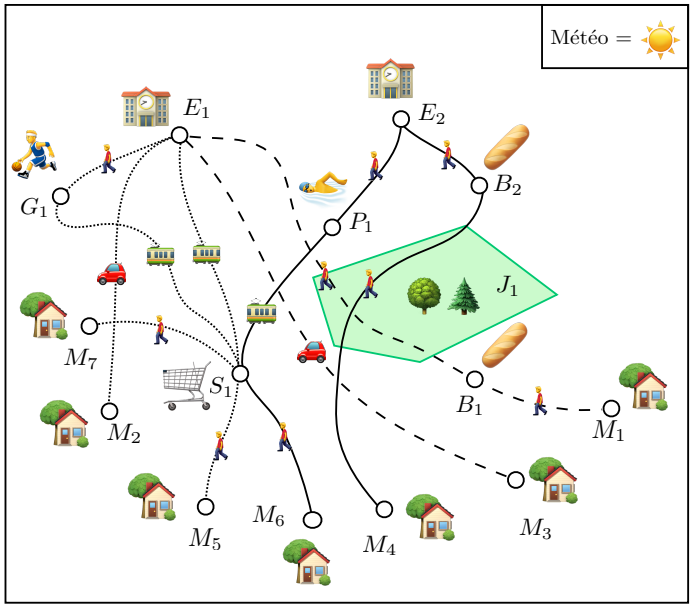

FIGURE *3. Exemple de 7 trajectoires sémantiques d'enfants*

cet exposé. Par convention, la trajectoire de l'enfant i se rendant à la maison i sera appelée trajectoire sémantique i (TS_i). Dans cet exemple, les enfants partent de deux écoles (E_1 et E_2) et retournent chez eux à pied, en voiture ou en tramway. Ici, uniquement les trajets retour les jours de beau temps ont été sélectionnés. La trajectoire TS_5 retrace le retour de l'enfant 5 qui part de l'école E_2 à pied pour se rendre à la piscine P_1 où il nage. Puis, il prend le tramway pour aller faire des courses au supermarché S_1 et finalement rentre à pied chez lui. Afin de faciliter la compréhension, la dimension temporelle n'a pas été prise en compte dans cet exemple. De même, seul le lieu est défini pour les activités statiques ou la trajectoire pour les activités mobiles. TS_5 est formalisée de la manière suivante :
$TS_5 = \langle$(Apprendre, E_2),(marcher, $TS_5.1$, à pied), (nager, P_1), (prendre tramway, $TS_5.2$, en tramway), (faire les courses, S_1), (marcher, $TS_5.3$, à pied), (`NULL`,M_5)$\rangle$ Les activités pratiquées à la maison étant inconnues, la dernière activité de cette séquence est `NULL`.

3. Mesures de comparaison et paramétrisation

3.1. Mesures de proximité classiques

Les contraintes utilisateurs et contextuelles étant des variables imprédictibles, il est nécessaire de constituer une mesure suffisamment adaptative et générique qui puisse correspondre avec précision à l'idée de similarité que porte l'utilisateur. Aussi, il subsiste un vide au sein des métriques existantes dans les SIG actuels et qui ne parvient pas à être comblé, une mesure de proximité générique qui puisse faire cohabiter les dimensions temporelle, géométrique et sémantique tout en sachant s'accorder avec les besoins utilisateurs.
Un panorama des mesures de similarité est dressé par (Li, 2014). Du côté géométrique, les distances à représentation linéaire, comme par exemple Dynamic Time Warping (DTW) (Sakoe, Chiba, 1978) et la distance de Fréchet (Devogele, 2002) sont très utilisées dans un contexte d'alignement des données et robustes pour les trajectoires possédant des sinuosités ou boucles. En contrepartie, elles s'appuient sur le calcul d'une matrice de distance, ce qui donne une complexité moyenne en $O(n \times m)$, où n et m correspondent aux nombres de points des trajectoires.
Le pendant sémantique est quant à lui représenté par deux familles de métriques : la distance d'édition et ses variantes - telles que Edit Distance on Real sequence (EDR) (Chen *et al.*, 2005) - basées sur la modification d'une suite de symboles et celles, comme Longest Common Subsequence (LCS), basées sur la recherche de la partie commune contiguë plus longue. Au sujet de la complexité ces algorithmes, (Wagner, Fisher, 1994) annonce une complexité moyenne en $O(n \times m)$. On notera que la distance d'édition est plus adaptative que LCS car elle offre la possibilité de considérer des opérateurs supplémentaires que ceux classiquement proposés (insertion, suppression, modification), de définir les coûts d'opération et tient compte de la totalité de la séquence de trajectoire. Pour ces raisons, nous proposons une métrique sémantique basée sur EDR implémentant les opérations de : Suppression, Insertion, Modification ainsi que la Permutation, Scission et Rassemblement. Ces trois dernières opérations sont en accord avec (Furtado *et al.*, 2016) qui avance le fait que deux trajectoires qui visitent les mêmes lieux (même sémantique) mais dans un ordre différent, peuvent être similaires. Les problèmes combinatoires que posent ces opérations de réarrangement, ou bien d'alignement, au sein des séquences souvent rencontrés en bio-informatique et commentés chez (Fertin *et al.*, 2009 ; Marteau, 2009).
Enfin, il est important de préciser que toute entité qualitative doit être considérée au sein d'une ontologie ou hiérarchie de concepts afin de pouvoir être soumise à comparaison. Dans (Aime, 2011) de nombreuses métriques sont proposées afin d'établir la similarité entre deux concepts ce qui permet d'effectuer les opérations de remplacement avec moins de rigidité.
Pour finir, concernant l'agrégation de différentes mesures afin de considérer les aspects géographiques, (Xu, Da, 2003) présente différents opérateurs existants

et communément utilisés pour agréger une série de valeurs. Les plus classiques sont les opérateurs min, max, moyenne pondérée ; un ensemble des possibilités sur les sommes et moyennes est établi dans (Grabisch *et al.*, 2011). Des solutions pour effectuer la similarité de séquences multi-dimensionnelles sont aussi énoncées dans (Furtado *et al.*, 2016 ; Gibert *et al.*, 2013).

3.2. Mesure générique de similarité entre trajectoires sémantiques

Cette section est dédiée à la présentation de la distance d'édition enrichie permettant d'officier sur les trajectoires sémantiques. Elle présente entre autres les différents opérateurs considérés puis un exemple d'instanciation de la mesure. L'aspect temporel n'est pas abordé, on suppose néanmoins qu'il est possible d'exercer sur les séquences d'activités un alignement temporel par un raisonnement selon l'algèbre des intervalles d'Allen, d'un algorithme d'alignement comme DTW et les opérateurs d'insertion, suppression.

Définition 1. — *Opérateurs d'édition*
Soient deux trajectoires sémantiques TS_1, TS_2 et un alphabet d'activités sémantiques Σ tel que $TS_1 \in \Sigma^n$ et $TS_2 \in \Sigma^p$. On rappelle que ε désigne en théorie des automates le symbole vide. Soient $TS_1 = \langle a_1, a_2, ..., a_n \rangle$ et $TS_2 = \langle a_i, a_j, ..., a_k \rangle$. Soient $a, b \in \Sigma$, tels que $a \neq b$ et le couple $(a, b) \neq (\varepsilon, \varepsilon)$. On considère l'ensemble d'opérations d'édition $E = \{\circledcirc, \otimes, \oplus, \ominus, \odot, \oslash\}$ tel que :

$$\forall e \in E \backslash \{\ominus\}, e : \begin{cases} \mathcal{TS} \times \mathbb{N} \times \Sigma \times \Sigma & \rightarrow \mathcal{TS} \\ (TS, k, a, b) & \mapsto e(TS, k, a, b) \end{cases}$$

De plus, on donne la signification suivante de $e(TS, k, a, b)$: "On applique l'opérateur e en remplaçant le symbole a à la position k dans TS par le symbole b". On définit les opérations de E telles que :

- *Modification $\oplus$: $a \rightarrow b$*
- *Insertion $\circledcirc$: $\varepsilon \rightarrow b$*
- *Suppression $\otimes$: $a \rightarrow \varepsilon$*
- *Permutation $\ominus$: On admet que $n = p$. On définit $\ominus : \mathcal{TS} \rightarrow \mathcal{TS}$ bijective telle qu'elle représente la permutation $\sigma = \begin{pmatrix} a_1 & a_2 & ... & a_n \\ a_i & a_j & ... & a_k \end{pmatrix} = \begin{pmatrix} TS_1 \\ TS_2 \end{pmatrix}$.*
- *Scission $\odot$: $(a \vee b) \rightarrow aba$*
- *Rassemblement $\oslash$: $aba \rightarrow (a \vee b)$*

On notera que $\forall e \in E, \exists e^{-1} \in E$ ce qui assure la symétrie .

Une séquence d'opérations d'édition transformant la trajectoire TS_1 en TS_2 est appelée chemin d'édition de TS_1 à TS_2, et $c(TS_1, TS_2)$ désigne l'ensemble des chemins d'édition de TS_1 à TS_2. Pour mesurer l'impact d'une opération d'édition sur la structure représentant un motif, nous définissons une fonction $\gamma : E \rightarrow \mathbb{R}^+$ qui assigne un coût d'édition à l'opération d'édition e. Les

faibles coûts correspondent à des opérations affectant peu la séquence selon l'utilisateur et les coûts élevés à des opérations jugées fortes. Le coût d'un chemin d'édition peut alors être déterminé par la somme de ses coûts d'opération d'édition individuels.

Définition 2. — *Distance d'édition sémantique*
La distance d'édition $d_{\mathcal{S}} : \mathcal{TS} \times \mathcal{TS} \to \mathbb{R}^+$ *, compte tenu de la fonction de coût d'édition* γ*, est le coût minimal pour transformer* TS_1 *en* TS_2*, soit :*

$$d_{\mathcal{S}}(TS_1, TS_2) = \min_{(e_1 \ldots, e_N) \in c(TS_1, TS_2)} \sum_{i=1}^{N} \gamma(e_i) \tag{1}$$

Soient deux trajectoires sémantiques TS_1 et TS_2 telles que :

– $TS_1 = \langle$(Apprendre, E_2), (se déplacer, $TS_1.1$, voiture), (nager, P_1), (se déplacer, $TS_1.2$, voiture), (faire les courses, S_1), (se déplacer, $TS_5.3$, voiture), (NULL,M_1)$\rangle$

– $TS_2 = \langle$(Apprendre, E_1), (se déplacer, $TS_2.1$, à pied), (basket-ball, G_1), (se déplacer, $TS_2.2$, voiture), (NULL,M_2)$\rangle$

Dans un but de concision, nous ne détaillerons pas les mesures utilisées pour l'opérateur de modification. On admet qu'il existe des mesures sémantiques afin calculer la proximité de deux concepts (Leacook, Chodorow, 1998) ou de deux instances disposant d'un ensemble de propriétés (Dice, 1945).
On peut définir la fonction de coût γ telle que :

$$\gamma(e) = \begin{cases} 1 & \text{si } e = \circledcirc \text{ ou } e = \otimes \\ 1 - \mathrm{Sim}(a, b) & \text{si } e = \oplus \\ 2\alpha \times \lg(\pi) & \text{si } e = \ominus \\ 2\beta & \text{si } e = \odot \text{ ou } e = \oslash \end{cases}$$

avec $(\alpha, \beta) \in [0, 1]^2, \mathrm{Sim} : \Sigma^2 \to [0, 1]$, la fonction calculant la similarité entre un couple d'activités sémantiques, lg la fonction retournant le nombre de transpositions de π issues de σ. On va transformer TS_1 en TS_2. On précise que les sous-trajectoires $TS_{\{1,2\}}.k$ ne sont pas considérées ici mais sont pris en compte dans le calcul d'une distance géométrique $d_{\mathcal{G}}(TS_1, TS_2)$.
On donne : $\alpha = 0.5, \beta = 0.5$, Sim((Apprendre, E_2),(Apprendre, E_1)) = 0.85, Sim((se déplacer, voiture), (se déplacer, à pied)) = 0.1, Sim((nager, P_1),(basket-ball, G_1)) = 0.65 et Sim((NULL, M_1), (NULL, M_2)) = 1.
Le chemin d'opérateurs $c(TS_1, TS_2)$ minimisant le coût de transformation de TS_1 vers TS_2 est : ($\oslash(TS_1$, 4, (se déplacer, voiture), (faire les courses, S_1)), $\oplus(TS_1$, 0, (apprendre, E_2), (apprendre, E_1)), $\oplus(TS_1$, 1, (se déplacer, voiture), (se déplacer, à pied)), $\oplus(TS_1$, 2, (nager, P_1), (basket-ball, G_1))).
Dès lors $d_{\mathcal{S}}(TS_1, TS_2) = 1 + 0.15 + 0.9 + 0.35 = 2.4$.
Enfin, considérant une distance géométrique $d_{\mathcal{G}}$ (DTW, Fréchet, ...), la distance

$d_{\mathcal{S}}$ de l'équation (1) et un opérateur d'agrégation Agg, il est possible de définir notre mesure de proximité $d : \mathcal{TS} \times \mathcal{TS} \rightarrow \mathbb{R}^+$ telle que :

$$d(TS_1, TS_2) = Agg(d_{\mathcal{S}}(TS_1, TS_2), d_{\mathcal{G}}(TS_1, TS_2)) \quad (2)$$

Supposons alors que $d_{\mathcal{G}}(TS_1, TS_2) = 3.5$. Soit l'opérateur d'agrégation moyenne pondérée de dimension 2 : $Agg(x, y) = \alpha x + (1 - \alpha)y$ avec $\alpha \in [0, 1]$. On pose ici $\alpha = 0.7$ pour donner peu plus de poids à la sémantique. Ainsi, l'équation (2) nous donne $d(TS_1, TS_2) = 0.7 \times 2.4 + 0.3 \times 3.5 = 2.73$.

4. Recherche, partitionnement et synthèse de motifs

4.1. Partitionnement des trajectoires sémantiques

Il existe peu de références à notre connaissance sur le partitionnement des trajectoires sémantiques, ceci dû principalement à l'absence d'une métrique pouvant réunir convenablement les dimensions temporelle, spatiale et sémantique. Ainsi, au sein de la fouille des trajectoires, l'aspect par partitionnement fût longtemps envisagé selon le prisme géométrique (Gianotti *et al.*, 2011).
Depuis peu, de nouvelles méthodes explorent l'angle sémantique et (Gibert *et al.*, 2013) commente l'apport d'éléments sémantiques et la prise en compte d'ontologies pour les clustering hiérarchiques (par partitionnement). (Ying *et al.*, 2014) s'approprie les différentes dimensions des trajectoires sémantiques mêlant ainsi partitionnement et fouille de motifs dans un dessein de prédiction. Les auteurs proposent une approche (GTS) basée sur les intentions des utilisateurs selon le contexte géographique, temporel et sémantique pour estimer la probabilité que l'utilisateur visite un lieu. L'idée centrale tient alors dans le calcul d'une similarité entre le mouvement actuel d'un utilisateur et les modèles GTS préalablement découverts.
Cependant, le partitionnement réalisé demeure très dépendant de la modélisation de la trajectoire sémantique et du type de mesure. Par exemple, (Xiao *et al.*, 2014) propose une mesure de similarité sémantique et une approche par partitionnement hiérarchique. Selon les auteurs, deux trajectoires sont considérées comme similaires si elles visitent la même séquence de lieux, plusieurs fois et avec un temps de déplacement similaire. Les permutations sont interdites.
Dans une veine similaire que celle réalisée par (Güting *et al.*, 2005), soit selon une méthode de représentation par la description de la position de l'objet par référencement linéaire à l'intérieur d'un réseau d'objets spatiaux en relation, (Wu *et al.*, 2015) propose une métrique selon le triptyque habituel au sein des réseaux routiers en tenant compte également de contraintes temporelles telles que l'horodatage (CTCP).
Enfin, en accord avec les techniques d'extraction d'information issue des médias sociaux et une modélisation de la trajectoire basée sur des régions d'intérêt, (Cai *et al.*, 2016) propose une méthode de partitionnement basée sur la densité.

Dans le cadre de la mesure proposée section 3.2, nous soutenons une approche par partitionnement hiérarchique. Si le partitionnement des trajectoires présentées figure 3. est effectué manuellement, deux configurations extrêmes possibles se dégagent : la première est celle où le paramètre géométrique est majoritairement valorisé. Dans ce cas, on observe des partitions figurées par les différents styles de pointillés représentées sur la figure 3., soient $C_1 = \{TS_1, TS_3\}, C_2 = \{TS_4, TS_6\}$ et $C_3 = \{TS_2, TS_5, TS_7\}$. Dans l'autre cas, celui où la sémantique prend le dessus, on observe un ensemble de partitions tel que $C_1 = \{TS_1, TS_4\}, C_2 = \{TS_2, TS_3\}$ et $C_3 = \{TS_5, TS_6, TS_7\}$.

4.2. Définition de motifs

La compréhension des motifs qui animent la mobilité est indispensable pour de nombreux domaines. Dans (Gianotti *et al.*, 2007), une extension du paradigme du fouille de motifs séquentiels est proposée pour l'analyse les trajectoires d'objets en mouvement. Ce modèle se base sur la découverte de régions d'intérêt et calcule les motifs fréquents de déplacement entre ces régions d'intérêt en intégrant des contraintes spatiales et temporelles. (Zhang *et al.*, 2014) élargi la recherche de motifs de déplacements à des ensembles de points d'intérêt géographiquement compacts, sémantiquement cohérents et dont les transitions temporelles entre ensembles surgissent rapidement (selon un seuil temporel donné). Pour les modèles de mouvement qui sont localement fréquents et non nécessairement dominants dans tout l'espace, (Choi *et al.*, 2017) s'inspire de la notion de compacité utilisée au sein de DBSCAN et l'adapte afin de quantifier la fréquence d'un motif particulier dans l'espace.
Ainsi, la recherche de motifs séquentiels utilise des méthodes à base de seuils (ou supports) qui parfois peuvent manquer de finesse en ne mesurant pas la ressemblance entre deux concepts. Une autre faiblesse est que, dans les modèles présentés, le lieu est considéré indépendamment de l'activité qui peut y être pratiquée ; Zhang et al., cependant, argumente le fait que, disposant des informations temporelles et géographiques, des activités peuvent être inférées à partir des médias sociaux. Des exemples allant dans ce sens sont donnés par (Long *et al.*, 2012 ; Yuan *et al.*, 2012) où les auteurs mettent en vedette une méthode par allocation de Dirichlet latente (LDA) permettant de déduire la fonction d'une région au sein d'une ville (par exemple des lieux d'enseignement, bureaux, zones de commerce) ou bien encore déterminer les relations intrinsèques et potentielles entre les lieux géographiques en utilisant les enregistrements de localisation qu'un utilisateur partage sur des médias sociaux basés sur la localisation. Convaincu d'une influence temporelle forte, (Zion, Lerner, 2017) étend le modèle LDA pour capturer l'influence du temps, en particulier de passé proche (jours/semaines), sur les motifs de mobilité des utilisateurs en utilisant des modèles temporels qui assouplissent les hypothèses de LDA afin de considérer au mieux les routines utilisateurs.

Les considérations précédentes montrent la difficulté réelle d'extraire les motifs de déplacement des trajectoires car bien souvent l'information demeure contextuelle. Aussi, si l'on se place dans le cadre d'une méthode par partitionnement préalable, un avantage est que, *a priori*, ces partitions formées offrent des ensembles de trajectoires homogènes aux comportements similaires. Il peut-être souhaitable, par la suite, d'en dresser une synthèse.
Cette vue synthétique peut être de nature géométrique ou sémantique selon les préférences de l'utilisateur. D'un point de vue géométrique, (Etienne *et al.*, 2016) propose le concept de trajectoire médiane sur l'appui de boîtes à moustaches spatio-temporelles. Le pendant sémantique peut quant à lui être assuré par une représentation sous forme d'automate ou de grammaire formelle (Mouza, Rigaux, 2005). En considérant un alphabet Σ de symboles sémantiques représentant les activités, il est alors possible à l'aide d'une inférence grammaticale sur une partition considérée d'extraire un langage la représentant en substance. Pour l'exemple de la figure 4. cette représentation synthétise un ensemble de déplacements qui partent de l'école à pied puis qui optionnellement vont faire une activité sportive. Un déplacement en tramway est alors réalisé pour aller faire des courses avant de rentrer à pied à la maison.

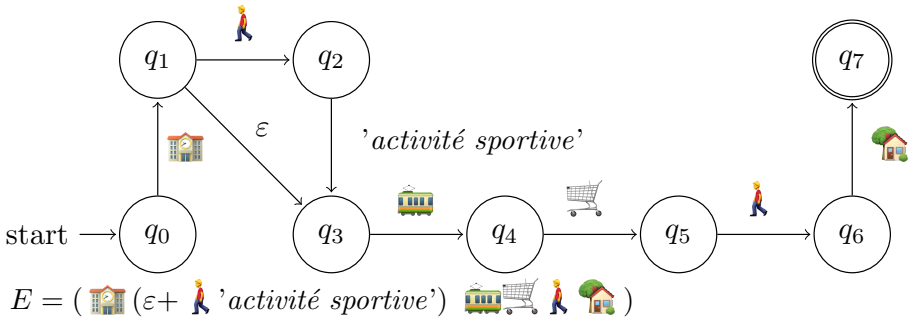

FIGURE *4. Automate et expression régulière représentant le motif synthétique de la partition de trajectoires* $\{TS_5, TS_6, TS_7\}$

5. Conclusions

Dans cet article a été présenté une méthode générique pour extraire les motifs de déplacements et d'activités des individus. La connaissance de ces motifs est fondamentale pour mieux comprendre les comportements humains. Ces travaux s'appuient sur une modélisation sémantiquement riche des activités des individus qui intègre les dimensions spatiale, temporelle et sémantique. Pour la dimension sémantique, plusieurs ontologies portant sur les lieux, les activités, les modes de déplacements et la météorologie sont intégrés. De même,

ces travaux nécessitent de réutiliser des distances spatiales ou temporelles et d'adapter des mesures de proximité symboliques telles que la distance d'édition. Finalement, ces travaux reprennent les outils de fouille de trajectoires qui extraient premièrement des partitions de trajectoires similaires ; pour chaque partition, un motif est inféré dans un second temps synthétisant les trajectoires sémantiques sous la forme d'un automate ou de grammaire. L'idée d'une grammaire probabiliste peut être envisagée afin de refléter le caractère stochastique des déplacements. Un avantage significatif de ces motifs est de résumer de manière anonyme un ensemble d'activités similaires. Dans le cas d'un nombre de trajectoires suffisant dans chaque groupe, cette synthèse peut répondre à des préoccupations éthiques légitimes pour l'analyse d'activités humaines.
De futurs travaux doivent être menés afin de rendre cette méthode générique et optimale. La dimension temporelle doit être approfondie, celle-ci pouvant être représentée sous un aspect linéaire ou cyclique, une approche générique doit être mise en place. De plus, dans cet article, plusieurs méthodes d'agrégation et mesures de similarité ont été rapidement présentées, elles doivent maintenant être mises en place et testées. Enfin, une solution adaptée pour la gestion simultanée de concepts ontologiques et de données spatio-temporelles volumineuses doit être proposée afin de manipuler de manière optimale ces données hétérogènes. Cet article forme ainsi un opuscule pour de futurs travaux où sera testée la méthodologie proposée sur les deux domaines d'application que sont la mobilité des enfants et les séjours touristiques.

Bibliographie

Aime X. (2011). *Gradients de prototypicalité, mesures de similarité et de proximité sémantique : une contribution à l'ingénierie des ontologies.* Thèse de doctorat, Université de Nantes.

Alvares L., Bogorny V., Kuijpers B., Macedo J. de, Moelans B., Vaisman A. (2007). A model for enriching trajectories with semantic geographical information. *Proc. of the 15th annual ACM international symposium on Advances GIS*, n° 22, p. 1–8.

Beber M., Ferrero C., Fileto R., Bogorny V. (2017). Individual and group activity recognition in moving object trajectories. *Journal of Information and Data Management*, vol. 8, n° 1, p. 50–66.

Bogorny V., Renso C., Aquino A. R. de, Lucca Siqueira F. de, Alvares L. (2014). Constant - a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, vol. 18, p. 66–88.

Cai G., Lee K., Lee I. (2016). Discovering common semantic trajectories from geo-tagged social media. In *Trends in applied knowledge-based systems and data science*, p. 320–332. Springer.

Chen L., Özsu M. T., Oria V. (2005). Robust and fast similarity search for moving object trajectories. *Proc. of the 2005 ACM SIGMOD*, p. 491–502.

Choi D., Pei J., Heinis T. (2017). Efficient mining of regional movement patterns in semantic trajectories. *Proc. of the VLDB*, vol. 10, p. 2073–2084.

Devogele T. (2002). A new merging process for data integration based on the discrete fréchet distance. In *Advances in spatial data handling*, p. 167–181. Springer.

Dice L. (1945). Measures of the amount of ecologic association between species. *Ecology*, vol. 26, p. 297–302.

Etienne L., Devogele T., Buchin M., McArdle G. (2016). Trajectory box plot; a new pattern to summarize movements. *International Journal of GIS*, vol. 30, p. 835–853.

Ferrero C., Alvares L., Bogorny V. (2016). Multiple aspect trajectory data analysis: Research challenges and opportunities. *GeoInfo*, p. 56–67.

Fertin G., Labarre A., Rusu I., Tannier E., Vialette S. (2009). *Combinatorics of genome rearrangements*. The MIT Press.

Furtado A., Kopanaki D., Alvares L., Bogorny V. (2016). Multidimensional similarity measuring for semantic trajectories. *Transactions in GIS*, vol. 20, p. 280–298.

Gianotti F., Nanni M., Pedreschi D., Pinelli F. (2007). Trajectory pattern mining. *ACM SIGKDD*, p. 330–339.

Gianotti F., Nanni M., Pedreschi D., Pinelli F., Rinzivillo S., Trasarti R. (2011). Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, vol. 20, p. 695–719.

Gibert K., Valls A., Batet M. (2013). Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering. *Knowledge and Information Systems*, vol. 40, p. 559–593.

González M., CA.Hidalgo, Barabási A.-L. (2008). Understanding individual human mobility patterns. *Nature*, vol. 453, p. 779–782.

Grabisch M., Marichal J.-L., Mesiar R., Pap E. (2011). Aggregation functions: Means. *Information Sciences*, vol. 181, p. 1–22.

Güting R., Almeida V. T. de, Ding Z. (2005). Modeling and querying moving objects in networks. *The VLDB Journal*, vol. 15, p. 165–190.

Leacook C., Chodorow M. (1998). Wordnet: An electronic lexical database. In, p. 265–283. Cambrige MA.

Li Z. (2014). Spatiotemporal pattern mining: Algorithms and applications. In, p. 283–306. Springer.

Long X., Lei J., Joshi . (2012). Exploring trajectory-driven local geographic topics in foursquare. *Proc. of the 2012 ACM Conference on Ubiquitous Computing*, p. 927–934.

Marteau P. (2009). Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, p. 306-318.

Mouza C. du, Rigaux P. (2005). Mobility patterns. *GeoInfo*, vol. 9, p. 297–319.

Noël D., Villanova-Oliver M., Gensel J., Quéau P. L. (2015). Modeling semantic trajectories including multiple viewpoints and explanatory factors: Application to life trajectories. *Proc. of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*, p. 107–113.

Parent C., Spaccapietra S., Renso C., Andrienko G., Bogorny V., Damiani M. *et al.* (2013). Semantic trajectories modeling and analysis. *ACM Computing Surveys*, vol. 45, p. 1–32.

Renso C., Trasarti R. (2013). Mobility data : Modeling, management and understanding. In, p. 129–151. Cambridge University Press.

Sakoe H., Chiba S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on ASSP*, vol. 26, p. 43–49.

Spaccapietra S., Parent C., Damiani M., Macedo J. de, Porto F., Vangenot C. (2008). A conceptual view on trajectories. *Data & Knowledge Engineering*, vol. 65, p. 126–146.

Wagner R., Fisher M. (1994). The string-to-string correction problem. *Journal of the ACM*, vol. 21, p. 168–173.

Wu X., Zhu Y., Xiong S., Peng Y., Peng Z. (2015). A new similarity measure between semantic trajectories based on road networks. *Proc. of the 17th Asia- Pacific Web Conference*, p. 522-535.

Xiao X., Zheng Y., Luo Q., Xie X. (2014). Inferring social ties between users with human location history. *Journal of AIHC*, vol. 5, p. 3–19.

Xu Z., Da Q. (2003). An overview of operators for aggregating information. *International Journal of intelligent systems*, vol. 18, p. 953–969.

Yan Z. (2009). Towards semantic trajectory data analysis: A conceptual and computational approach. *VLDB Endowment*.

Yan Z., Chakraborty D., Parent C., Spaccapietra S., Aberer K. (2011). Semitri: A framework for semantic annotation of heterogeneous trajectories. *Proc. of the 14th International Conference on Extending Database Technology*, p. 259–270.

Yan Z., Parent C., Spaccapietra S., Chakraborty D. (2010). A hybrid model and computing platform for spatio-semantic trajectories. *Proc. of the 7th international conference on The Semantic Web*, p. 60-75.

Ying J.-C., Lee W.-C., Weng T.-C., Tseng V. (2014). Mining geographic-temporal-semantic patterns in trajectories for location prediction. *ACM Transactions on Intelligent Systems and Technology*, vol. 5, n° 2, p. 1–33.

Yuan J., Yu Z., Xing X. (2012). Discovering regions of different functions in a city using human mobility and pois. *Proc. of the 18th ACM SIGKDD*, p. 186-194.

Zhang C., Han J., Shou L., Lu J., Porta T. L. (2014). Splitter: Mining fine-grained sequential patterns in semantic trajectories. *VLDB Endowment*, vol. 7, p. 769-780.

Zion E., Lerner B. (2017). Learning human behaviors and lifestyle by capturing temporal relations in mobility patterns. *Proc., European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, p. 459–464.