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 à l'aide d'une ontologie 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 constitue une étude bibliographique et présente une méthodologie afin d'extraire ces motifs à partir de trajectoires sémantiquement riches. 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 using an ontology. 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 bibliography study and a new methodology about how 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, Ontologie, Fouille de trajectoires*

KEYWORDS: *Semantic trajectory, Trajectory clustering, Moving Object patterns, Similarity measure, Ontology, 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 ; Song *et al.*, 2010) présente que malgré la complexité et la 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 les résultats précédemment énoncés mettent en lumière le caractère prédictible des trajectoires humaines, ils ne fournissent pas d'explication satisfaisante quant au sens et au contexte des déplacements (Renso, Trasarti, 2013 ; Yan, 2009 ; Parent *et al.*, 2013).
C'est au sein de cette brèche sémantique que nous posons 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.
Aussi, l'abondance de ressources complémentaires au GPS, telles que les ontologies ou les médias sociaux 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 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) au sein des SIG actuels, 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*) ? 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. Ainsi, elle se présente avant tout comme un opuscule bibliographique, illustré selon des cas d'école, sur les sujets liés à l'extraction de motifs dans le cadre des trajectoires enrichies sémantiques.
L'article est organisé comme suit : Dans la section 2 , plusieurs modèles de trajectoires sémantiques sont discutés et nous présentons un modèle sémantique adéquat afin de traiter les cas hypothétiques de travail 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, 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 avant de dresser nos conclusions.

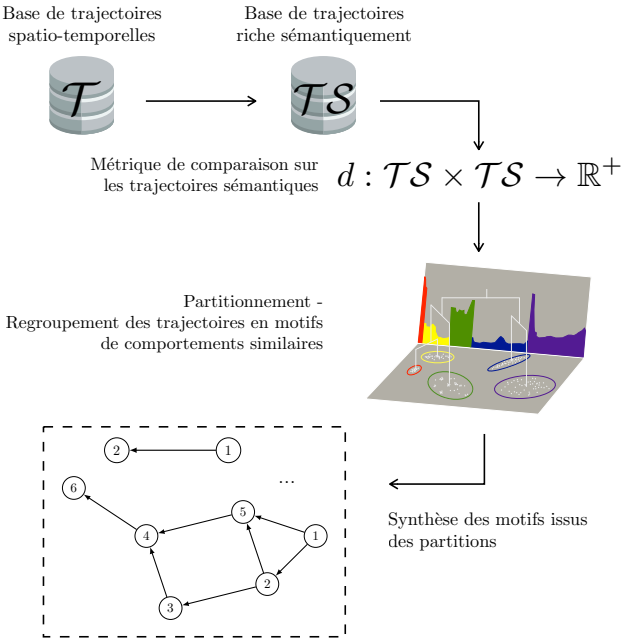

FIGURE *1. Chaîne de traitement pour l'extraction de motifs de trajectoires sémantiques similaires*

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. Par exemple dans (Yan *et al.*, 2010), la trajectoire sémantique est représentée comme une séquence d'épisodes sémantiques et 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 (Transports, horaires d'événements, d'activités, routines ou 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) : Une des améliorations significatives serait l'alimentation du contexte des trajectoires par des ontologies afin de mener des inférences sur les activités ou bien d'établir une hiérarchie de concepts qui permettrait alors de calculer plus finement la proximité sémantique de 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 ; Miller, Han, 2009) 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. Une **trajectoire sémantique** d'individu est une suite d'activités $TS = \{a_1, ..., a_n\}$ ordonnée temporellement. Une **activité** a est décrite selon les trois dimensions : sémantique, spatiale et temporelle. La dimension sémantique est apportée à l'aide d'un ensemble d'ontologies $\mathcal{O}$. Les concepts ontologiques sont organisés hiérarchiquement par des liens is_a selon un arbre ou un treillis ce qui permet de définir une notion de proximité entre concepts (Aime, 2011). Par exemple les concepts "Natation" et "Basket-ball" sont des sous-concepts du concept "Acti-

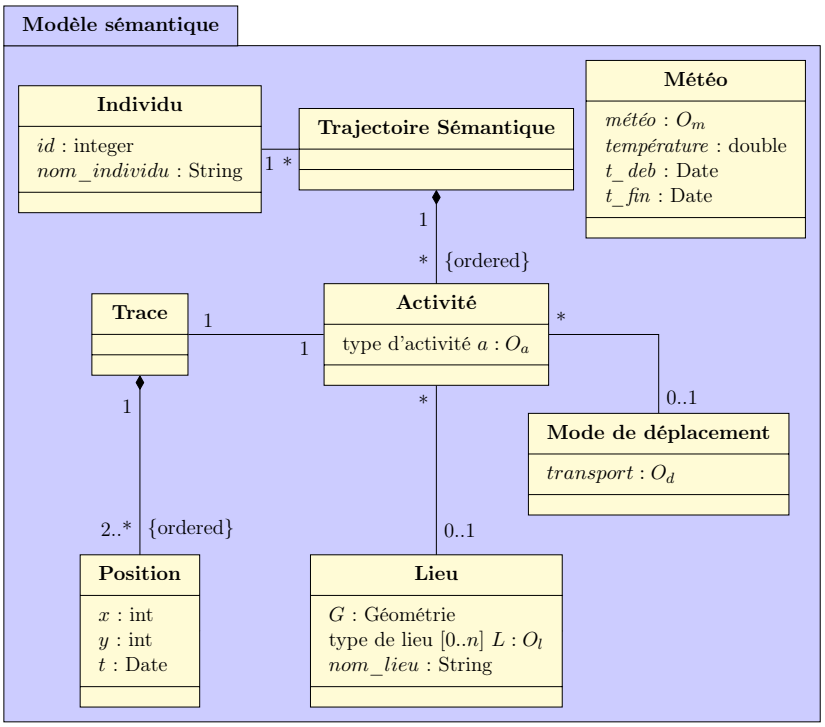

FIGURE *2. Modélisation UML des trajectoires sémantiques d'individus*

vité sportive". Ainsi, dans le cas des activités, tout type d'activité $a \in O_a$ avec $O_a \in \mathcal{O}$, c'est-à-dire que tout concept d'activité doit appartenir à l'ontologie. À noter que le concept `Null` est autorisé, il représente une activité inconnue ou une activité d'attente.

Les activités sont associées à des **traces** [brutes] (ou *raw data*). Ces objets spatio-temporels regroupent des **positions** ordonnées temporellement sous la forme de suites de triplets $p_i = (x, y, t)$. Une activité peut-être statique ou mobile, le distinguo est fait si l'on constate que la trace $T = \{p_1, ..., p_k\}$ associée à l'activité a est de la forme $\exists (x, y) \in \mathbb{R}^2, \forall p \in T, (p.x, p.y) = (x, y)$, c'est-à-dire que la trace se concentre en un point fixe géographique. De même, c'est la trace qui porte la dimension temporelle et qui délimite la durée d'activité d_a, soit $d_a = [p_1.t, p_k.t]$.

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 de mobilité et représente alors une activité mobile.

Un dernier point à signifier est que la trajectoire sémantique est organisée de

manière à éviter tout recouvrement temporel. Ainsi, $\forall a_i, a_j \in TS, d_{a_i} \cap d_{a_j} = \emptyset$; autrement dit, il n'est pas permis d'effectuer plus d'une activité simultanément, ce qui complexifierait le modèle de manière importante ; il peut néanmoins être défendu que, dans bien des cas d'activités simultanées, il existe une activité qui prévaut par rapport aux autres. Cependant, si l'on considère la trajectoire sémantique TS non plus comme plus une suite d'activités ordonnées temporellement mais comme une suite d'ensemble d'activités de la forme $TS = \{A_1, ..., A_n\}$ où $A_i = \{a_1, ..., a_p\}$, alors il est possible de tenir compte d'activités simultanées. Cependant, on perd l'ordre temporel total établie entre les activités ; les raisonnements temporels peuvent néanmoins être menés l'aide des règles de l'algèbre des intervalles d'Allen.

Les activités sont potentiellement attachées à des **lieux** qui sont des instances du monde physique. Un lieu est décrit par une géométrie G telle que $a.T \subseteq G$, c'est-à-dire que l'on considère le lieu dont la géométrie englobe spatialement l'activité, et un ensemble de concepts de type de lieux $L \subseteq O_l$ dont les éléments $l \in L$ sont les parents directs de l'instance décrite. Par exemple, le lieu P_1 est une instance attachée au concept parent $L_{P_1} = \{\text{piscine}\}$ et G_1 une instance du concept parent $L_{G_1} = \{\text{gymnase}\}$. Ces instances sont reliées entre-eux par le concept plus général "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 météorologique $m \in O_m$.

Les 7 trajectoires schématiques de la FIGURE *3.* vont servir d'exemple fil

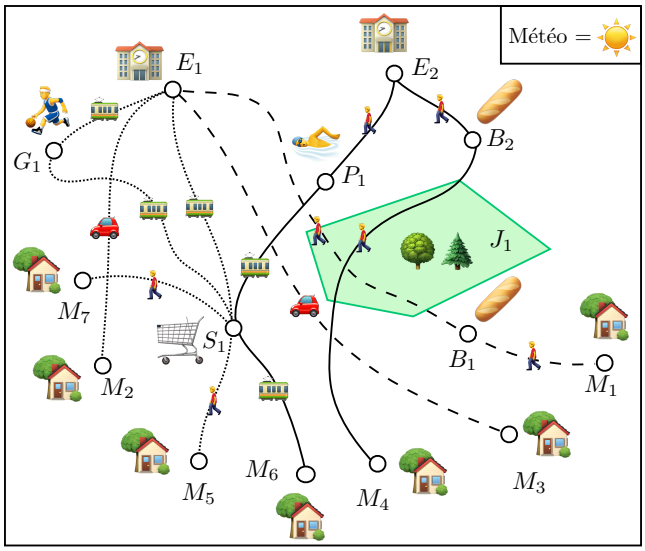

FIGURE *3. Exemple de 7 trajectoires sémantiques d'enfants*

rouge à cet exposé. Par convention, la trajectoire sémantique 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 de concept parent respectif L_{E_1} = {école public} et L_{E_2} = {école privé}) et retournent chez eux à pied, en voiture ou en tramway. On considère que la météo est fixée à "Beau temps" lors de l'analyse.

Exemple. — Modélisation de la trajectoire sémantique TS_6
La trajectoire sémantique TS_6 retrace le retour de l'enfant 6 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 en tramway chez lui. Afin de faciliter la compréhension, la dimension temporelle n'a pas été prise en compte dans cet exemple. De même pour les activités statiques, seul le lieu est défini (la trace se réduisant à un point) : $TS_6 = \langle$(apprendre, E_2),(marcher, $TS_6.1$, à pied), (nager, P_1), (se déplacer, $TS_6.3$, en tramway), (faire les courses, S_1), (se déplacer, $TS_6.5$, en tramway), (`Null`,M_6)$\rangle$.
Où $TS_i.k$ est la trace de l'activité a_k, elle n'est indiquée que dans le cadre d'une activité mobile (une activité fixe se réduisant spatialement à un point). □

Les activités pratiquées à la maison étant inconnues, la dernière activité de cette séquence est `Null`.

3. Mesures de comparaison et paramétrage

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 ou encore qui lui permette de s'adapter selon ces besoins. 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 d'une matrice de distance d'où une complexité moyenne de calcul 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 la distance d'édition et implémentant les opérations de : Ajout, Suppression, Modification ainsi que l'insertion et la permutation. Cette dernière opérations est 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 liés à ces opérations de réarrangement ou 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 de la section 2. 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 algorithme comme DTW.

Définition 1. — *Opérateurs d'édition*
On définit $\Sigma = \{(c_a, c) | c_a \in O_a, c \in (O_d \cup O_l)\}$ *un alphabet d'activités. Soient deux trajectoires sémantiques* TS_1, TS_2 *telles que* $TS_1 \in \Sigma^n$ *et* $TS_2 \in \Sigma^p$. *On rappelle que* ε *désigne le symbole vide et peut-être rattaché au concept* `Null`. *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* $E = \{\texttt{add}, \texttt{supp}, \texttt{modif}\}$ *(ajout, suppression, modification) où* $\forall e \in E, e : \Sigma \times \Sigma \rightarrow \Sigma$. *On définit également l'opérateur d'insertion* $\texttt{insert} : \Sigma \times \Sigma \rightarrow \Sigma^3$ *et l'opérateur* $\texttt{sort} : \Sigma^n \times \Sigma^n \rightarrow \Sigma^n$ *de réarrangement.*

Ces opérateurs sont définis tels que :

– `modif` : $a \to b$ *(Remplacement de l'activité a par b).*

– `add` : $\varepsilon \to b$ *(Ajout de l'activité b).*

– `supp` : $a \to \varepsilon$ *(Suppression de l'activité a).*

– `insert` : $a \to aba$ *(Insertion d'une activité b durant l'activité a). Cet opérateur peut-être utile notamment dans un contexte Move-Stop-Move, pour marquer une course durant un déplacement par exemple. On considérera aussi l'opérateur inverse* $\texttt{insert}^{-1} : aba \to a$.

– *On admet que* $n = p$ *et que* $\forall a \in TS_1, a \in TS_2$ *et* $\forall a \in TS_2, a \in TS_1$, *alors* `sort` *représente la permutation* $\sigma = \begin{pmatrix} a_1 & a_2 & ... & a_n \\ a_i & a_j & ... & a_k \end{pmatrix} = \begin{pmatrix} TS_1 \\ TS_2 \end{pmatrix}$. *On représente également* σ *selon sa décomposition canonique en permutations circulaires telle que* $\sigma = \pi_1 \circ ... \circ \pi_p$. *On notera* $\lg(\sigma) = p$.

De plus, on définit la fonction de coût $\gamma : E \to \mathbb{R}^+$ *qui associe à un opérateur son prix d'utilisation. Il est possible d'ajouter de nouveaux opérateurs, mais il faut veiller à ce que* $\forall e \in E, \exists e^{-1} \in E$ *afin de garantir la symétrie .*

Une séquence d'opérations d'édition $(e_1, e_2, ..., e_N)$ transformant la trajectoire TS_1 en TS_2 est appelée un chemin d'édition $c(TS_1, TS_2)$ et correspond à la composition successive des opérations e_1 à e_N. On désigne par $\mathcal{C}(TS_1, TS_2)$ l'ensemble de tous les chemins d'édition de TS_1 à TS_2. 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 $d_{\mathcal{S}} : \Sigma^p \times \Sigma^n \to \mathbb{R}^+$ *, compte tenu de la fonction de coût d'édition* γ*, est le coût minimal pour transformer* TS_1 *en* TS_2*, soit :*

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

Exemple. — Instanciation de γ
Soient les trajectoires sémantiques TS_6 et TS_5 de la Figure *3*, on détaille TS_5 telle que : $TS_5 = \langle$(apprendre, E_1), (se déplacer, $TS_5.1$, en tramway), (basket-ball, G_1), (se déplacer, $TS_5.3$, en tramway), (faire les courses, S_1), (se déplacer, $TS_5.5$, à pied), (`Null`,M_6)$\rangle$.
On peut définir la fonction de coût γ telle que :

$$\gamma(e) = \begin{cases} 1 & \text{si } e = \texttt{add} \text{ ou } e = \texttt{del} \\ 1 - \mathrm{Sim}(a, b) & \text{si } e = \texttt{mod} \\ 2\alpha & \text{si } e = \texttt{insert} \text{ ou } e = \texttt{insert}^{-1} \\ \beta \sum\limits_{\pi_i \in \sigma} \sum\limits_{k=1}^{\lg(\pi_i)-1} \mathrm{Sim}\left(\pi_i(a_k), \pi_i(a_{k+1})\right) & \text{si } e = \texttt{sort} \end{cases}$$

où Sim : $\Sigma \times \Sigma \rightarrow [0,1]$ est une fonction calculant la similarité entre un couple d'activités, on notera que Sim doit être symétrique, c'est-à-dire que $\text{Sim}(a,b) = \text{Sim}(b,a)$ [1] ; et $(\alpha, \beta) \in]0,1]^2$, on remarquera que si $\alpha = 1$ (resp. $\beta = 1$), alors l'opération associée est équivalente à un traitement selon les opérateurs classiques `add`, `supp`, `modif`. Par exemple, pour $\alpha = 1$, l'opérateur d'insertion est équivalent à l'application successive de deux ajouts. Une bonne pratique étant de borner le coût des opérateurs entre 0 et 1 pour la modification d'un symbole.
On va transformer TS_5 en TS_6. On précise que les traces $TS_{\{5,6\}}.k$ ne sont pas considérées ici mais sont pris en compte dans le calcul d'une distance géométrique $d_{\mathcal{G}}$.
On donne : $\alpha = 0.5, \beta = 0.5$. Un tel paramétrage permet de réduire de moitié le coût des opérations de permutation et insertion, c'est-à-dire comparativement au cas où elles auraient été effectuées à l'aide des opérateurs classiques. On suppose que la fonction de similarité donne les valeurs suivantes pour les couples considérés : Sim((apprendre, E_2),(apprendre, E_1)) = 0.85, Sim((se déplacer, en tramway), (se déplacer, à pied)) = 0.7, Sim((nager, P_1),(basket-ball, G_1)) = 0.65 et Sim((`NULL`, M_1), (`NULL`, M_2)) = 1.
Le chemin d'opérateurs $c(TS_5, TS_6)$ minimisant $d_{\mathcal{S}}$ est :

1. $\texttt{modif}(TS_5^{(0)}, TS_6^{(0)}) = \texttt{modif}((\text{apprendre}, E_2), (\text{apprendre}, E_1))$,
2. $\texttt{modif}(TS_5^{(2)}, TS_6^{(2)}) = \texttt{modif}((\text{nager}, P_1), (\text{basket-ball}, G_1))$,
3. $\texttt{sort}(TS_5, TS_6)$. On a ici $\sigma = \pi_1 = (TS_5^{(1)}, TS_5^{(5)})$

L'application de $\texttt{sort}(TS_5, TS_6)$ en toute fin du chemin d'édition permet de permuter les activités $TS_5^{(1)}$ et $TS_5^{(5)}$.
Dès lors $d_{\mathcal{S}}(TS_5, TS_6) = \gamma(\texttt{modif}(TS_5^{(0)}, TS_6^{(0)})) + \gamma(\texttt{modif}(TS_5^{(2)}, TS_6^{(2)})) + \gamma(\texttt{sort}(TS_5, TS_6)) = (1 - 0.85) + (1 - 0.65) + \frac{1}{2} \times (1 - 0.7) = 0.65$.
□

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 :

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

Supposons alors que $d_{\mathcal{G}}(TS_1, TS_2) = 3.5$. On pose l'opérateur d'agrégation moyenne pondérée de dimension 2 : $Agg(x, y) = \alpha x + (1 - \alpha)y$ avec $\alpha \in [0,1]$. Une telle fonction d'agrégation permet selon le paramètre α de privilégier la dimension sémantique ou géographique, on posera entre autre $\alpha = 0.5$ si l'on souhaite traiter ces deux dimension de manière égale.
Supposons que l'on pose $\alpha = 0.7$ afin de privilégier l'aspect sémantique. Ainsi, l'équation (2) nous donne $d(TS_1, TS_2) = 0.7 \times 0.65 + 0.3 \times 3.5 = 1.505$.

1. De nombreux exemples de mesures de similarité entre concepts sont exposées dans (Aime, 2011 ; Leacook, Chodorow, 1998 ; Dice, 1945).

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 méthodes de clustering hiérarchique (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

On a vu en section 2 que de nombreux modèles de représentation de la trajectoire sémantique utilisé des modes de représentation sous forme de séquence. Dans cette lignée, (Gianotti *et al.*, 2007) propose une extension du paradigme de la fouille de motifs séquentiels pour l'analyse les trajectoires d'objets en mouvement. Les auteurs représentent les trajectoires selon le paradigme *Stop and Move* comme des séquences de régions d'intérêt visitées séparées par un temps de déplacement. Le modèle de fouille se base alors sur la découverte de régions d'intérêt en calculant les motifs fréquents de déplacement entre ces régions d'intérêt sous contraintes de seuils spatiaux et temporels. (Zhang *et al.*, 2014) élargit 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 encore de déterminer les relations intrinsèques et potentielles entre les lieux géographiques en utilisant les enregistrements de localisation qu'un utilisateur partage sur des médias sociaux. 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 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 $\{TS_5, TS_6, TS_7\}$ des enfants qui partent de l'école à pied ou en tramway et qui vont potentiellement faire une activité sportive. Ces enfants prennent ensuite le tramway pour aller faire des courses avant de rentrer à pied ou en tramway chez eux.

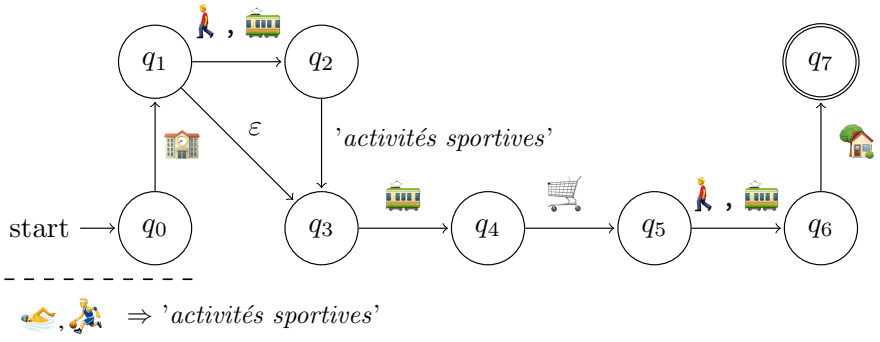

FIGURE *4. Automate et inférence logique représentant le motif synthétique de la partition de trajectoires* $\{TS_5, TS_6, TS_7\}$

5. Conclusions

Dans cet article a été présentée 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 intégrées. De même, ces travaux nécessitent de réutiliser des distances spatiaux-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 ma-

nière 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.
Ainsi cet article, une nouvelle distance d'édition pour la similarité entre trajectoires sémantiques afin d'extraire des motifs de comportement semblables. De futurs travaux doivent être menés afin de rendre cette méthode générique et optimale, en particulier la dimension temporelle doit être approfondie ainsi que les méthodes de similarité entre symboles d'un alphabet afin de rendre complète la métrique, enfin, la distance pourra être mise en œuvre et testée. Également, une solution adaptée pour la gestion simultanée de concepts ontologiques et de données spatio-temporelles volumineuses devra être proposée afin de manipuler ces données hétérogènes de manière optimale.
Cet article forme ainsi un opuscule pour de futurs travaux où sera testée la méthodologie proposée sur des données réelles issues des deux domaines d'application présentés en introduction que sont la mobilité des enfants et les séjours touristiques.

6. Remerciements

Ce travail a été financé et soutenu par l'ANR via le projet Mobikids, ainsi que par la région Centre-Val de Loire via le projet de recherche d'intérêt régional Smart Loire.

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.

Miller H., Han J. (2009). *Geographic data mining and knowledge discovery*. Taylor & Francis.

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.

Song C., Qu Z., Blumm N., Barabási A.-L. (2010). Limits of predictability in human mobility. *Science*, vol. 327, p. 1018–1021.

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