Une méthode de découverte de motifs contextualisés dans les traces de mobilité d'une personne

Aimene Belfodil*, Mehdi Kaytoue*
Céline Robardet*, Marc Plantevit**, Julien Zarka***

*Université de Lyon, CNRS, INSA-Lyon, LIRIS UMR5205, F-69621, Lyon, France
**Université de Lyon, CNRS, Univ. Lyon 1, LIRIS UMR5205, F-69622, Lyon, France
***Mobile Devices, 100 Avenue de Stalingrad, F-94800, Villejuif, France
contact : aimene.belfodil@insa-lyon.fr

Résumé. Les traces de mobilité générées par les divers capteurs qui nous entourent peuvent être analysées à des fins prédictives et explicatives pour répondre à divers problèmes du quotidien. Si de nombreuses méthodes ont été proposées pour décrire le comportement d'un individu de manière globale à partir des transitions entre ses différents points d'intérêts (par exemple via un modèle de Markov), peu de travaux cherchent à l'expliquer de manière locale. Nous proposons dans cet article une méthode qui permet d'extraire pour un individu dont on a une trace de mobilité conséquente des motifs de mobilité dits contextualisés. Chaque motif est composé d'une description sur l'ensemble des visites aux différents points d'intérêt de l'individu qui maximise une ou plusieurs mesures avec une sémantique particulière (le motif décrit une phase sédentaire ou exceptionnel de la mobilité de l'individu). Une expérimentation a été menée à partir de traces de mobilité de véhicules et donne des résultats encourageants.

1 Introduction

L'analyse de traces de mobilité suscite depuis plusieurs années un intérêt grandissant. En effet, des capteurs capable de nous géo-localiser sont omniprésents dans notre quotidien (téléphone mobile, smart-watch, ...). Ils produisent des traces de mobilité véloces et volumineuses. La disponibilité de ces traces rend possible un vaste champ d'applications afin, par exemple, d'améliorer la gestion du trafic urbain (Zhang et al., 2013) ou d'offrir un service de covoiturage (Trasarti et al., 2011). Une autre application, qui motive le présent travail, vise à identifier des profils de conduite afin, par exemple, d'adapter au plus près une police d'assurance aux risques encourus par l'usager : un individu peut effectuer des trajets exceptionnels, être sédentaire, etc.

Dans cet article, notre objectif est de décrire de manière automatique le comportement d'un objet mobile à l'aide de motifs. Plusieurs travaux ont été proposés pour décrire la mobilité d'un ou plusieurs objets mobiles. Parmi ces travaux, on peut citer la méthode de Gambs et al. (2011) qui construit un graphe de mobilité, ou celle de Gonzalez et al. (2008) qui repose sur une description statistique. Cependant, les motifs de mobilité mis à jour par les méthodes de la littérature sont souvent trop globaux, comme un modèle de Markov qui encode les probabilités de transitions entre les nœuds ou points d'intérêt (POI). Ces méthodes ne permettent pas

d'obtenir des motifs contextualisés c'est-à-dire munis d'une explication quant aux conditions dans lesquelles le motif se réalise.

Pour atteindre notre objectif, nous proposons une méthode de fouille de données en deux étapes. La première consiste à transformer la trace de mobilité, initialement une séquence de coordonnées spatiales estampillées, en une suite de visites aux positions fréquemment visitées (POI). Chaque visite est décrite par le POI visité, ainsi qu'une description qui inclut à minima l'intervalle de temps passé dans ce POI, mais qui peut être étendue à l'aide de descripteurs spatiaux (informations sur les POI) ou temporels (jour de la semaine, jours fériés, météorologiques, etc.). La deuxième étape consiste à générer des motifs sur l'espace de descriptions des visites et à ne retenir que les plus pertinents par rapport à une mesure de qualité. Plusieurs mesures de qualité peuvent être utilisées, chacune véhiculant une sémantique particulière (Sédentarité de l'individu, Exceptionnalité des visites ...).

Cette méthode d'extraction de motifs est ancrée dans (i) l'analyse de concepts formels (*formal concept analysis* ;Ganter et Wille (1999)) pour l'extraction de motifs et (ii) dans la découverte de sous-groupes (*subgroup discovery* ; Novak et al. (2009)). Outre l'aspect méthodologique, nos contributions sont les suivantes, (i) nous introduisons un nouveau domaine de motifs pour la fouille des traces de mobilité et (ii) nous montrons comment étendre un algorithme existant pour considérer des attributs avec un domaine de valeurs de type intervalle. Enfin (iii) nous validons notre approche à travers une expérimentation sur des données réelles.

La suite de cet article est organisée comme suit. La section 2 introduit les définitions préliminaires et positionne le problème. La section 3 introduit l'algorithme d'extraction de motifs de mobilité qui est expérimenté sur des données réelles en section 4 montrant l'intérêt de l'approche et ses premiers résultats.

2 Notations et problème

Nous considérons qu'à un utilisateur est associé une trace de mobilité. Une trace de mobilité pour un objet O est ainsi une liste ordonnée d'évènements $P^{(O)} = \{p_1^{(O)}, p_2^{(O)}, ..., p_n^{(O)}\}$ où chaque événement est décrit *a minima* par une estampille temporelle et ses coordonnées spatiales définies par un vecteur à n dimensions. Par exemple, les événements de traces de mobilité issues de téléphones portables ou de boîtiers GPS sont décrits par la latitude et la longitude des objets correspondants à un instant donné. Les événements peuvent être également enrichis par d'autres attributs décrivant l'état de l'objet à l'instant de l'évènement (vitesse, accélération, niveau batterie, etc.). A partir d'une trace de mobilité brute, nous pouvons définir les notions de *point d'intérêt*, *visite*, et de *base de visites*.

Un *point d'intérêt* (noté POI) d'un objet O est un lieu jugé pertinent pour cet objet (i.e., fréquemment visité). Il est décrit par ses coordonnées $(x_1, \ldots, x_n)$ et éventuellement d'autres descripteurs comme par exemple un polygone ou encore des descripteurs sémantiques le caractérisant (e.g., lieu touristique, parc d'attraction, commerce). Un objet O est associé à un ensemble de points d'intérêt noté $POI^{(O)}$.

Une *visite* v d'un objet O décrit la présence de l'objet sur un POI à un intervalle de temps donné $[t_A, t_D]$. Formellement, une visite v est donc un couple $(p_v, [t_A, t_D])$ où $p_v \in POI^{(O)}$ et t_A, t_D des instants temporels. Une visite peut être enrichie par d'autres attributs (e.g., des agrégats sur les événements de la visite comme le taux de précipitation, la température).

poi	t_A	t_D
$poi_1^{(O)}$	14/01/2015 21:00	15/01/2015 08:23
$poi_2^{(O)}$	15/01/2015 08:47	15/01/2015 11:49
$poi_3^{(O)}$	15/01/2015 12:35	15/01/2015 13:53
$poi_1^{(O)}$	15/01/2015 14:07	15/01/2015 18:17
$poi_1^{(O)}$	15/01/2015 18:42	15/01/2015 20:55

TAB. 1 – *Exemple - Base des visites*

On peut ainsi associer à un objet O une *base de visites* $V^{(O)}$ qui contient toutes ses visites (voir Table 1). Cette base $V^{(O)}$ décrites sur les attributs $A = \{a_1, a_2, ..., a_n\}$ est un ensemble sur lequel on peut calculer des *descriptions* ou *motifs* évalués par le biais d'une *mesure de qualité*. L'attribut a_i peut être de type symbolique, numérique ou intervalle numérique. Nous définissons ces deux notions.

Une *description d'un motif* (ou contexte) notée $d = (R_1, R_2, ..., R_n)$ est une liste de restrictions R_i sur le domaine de valeurs de chacun des attributs a_i. On dit qu'une visite $v \in V^{(O)}$ *vérifie la description* d ssi pour chacun des attributs a_i de v : (i) $a_i \in R_i$ si a_i est un attribut symbolique ou numérique, (ii) $a_i \subseteq R_i$ si l'attribut est un intervalle numérique, R_i est un intervalle numérique si a_i est un attribut de type numérique ou intervalle numérique, et un ensemble de symboles si a_i est un attribut symbolique. L'ensemble de visites qui *vérifient* la description d est appelé support, noté $supp(d)$. D désigne l'ensemble des descriptions sur $V^{(O)}$.

Une *mesure de qualité* $\varphi : D \rightarrow R$ est une mesure qui évalue la pertinence d'une description d'un motif d, et permet d'établir un ordre entre les motifs. Plusieurs mesures de qualité peuvent être utilisées, véhiculant chacune une sémantique particulière. On propose de mesurer la *sédentarité* de l'individu en utilisant la notion d'entropie. En effet plus l'entropie est faible, plus la description est caractéristique d'un nombre restreint de POI :

$$\varphi_{sed}(d) = \sum_{poi \in POI^{(O)}} -log_2(\beta_{poi}(d)) \times \beta_{poi}(d) \text{ avec } \beta_{poi}(d) = \frac{|\{v \in supp(d) | p_v = poi\}|}{|supp(d)|}$$

Problème 1 (Découverte de motifs contextualisés dans les traces de mobilité) *Étant donnée une trace de mobilité, transformée en une base de visites, on cherche les motifs ou contextes qui minimisent la mesure de sédentarité φ_{sed} et couverts par au moins $minSup$ visites.*

3 Un algorithme de découverte de motifs contextualisés

Pour découvrir les motifs de mobilité contextualisés, nous proposons un algorithme générique qui (i) énumère les contextes à partir de la *base des visites* de l'objet mobile et calcule la *mesure de qualité* pour chacun (ii) fait un post-traitement pour retenir uniquement les meilleurs selon certains critères. Puisque de nombreux contextes peuvent prendre un ensemble de visites comme image, nous basons notre algorithme sur l'énumération de contextes fermés (Ganter et Kuznetsov (2001)). Cette approche nécessite cependant que les mesures de qualité ne dépendent que du support ce qui est le cas pour la mesure de sédentarité. Pour gérer les données hétérogènes, où chaque attribut prend des valeurs soit numériques soit symboliques, un algorithme a déjà été proposé dans ce cadre par Kaytoue et al. (2011). Il énumère les motifs par spécialisation, ce que nous voulons, mais n'est pas adapté quand les objets peuvent prendre un intervalle de valeurs pour un attribut numérique (au lieu d'une unique valeur). Dans cette section, on montre son adaptation pour le cas des données intervalles, une de nos contributions.

1 - Énumération des contextes fermés. Soient G un ensemble d'objets (base de visites), $(D, \sqcap)$ un inf-demi-treillis où D désigne l'ensemble des descriptions possibles ordonnées

par la relation de subsomption $c \sqsubseteq d \Leftrightarrow c \sqcap d = d, \ \forall c, d \in D$, et $\delta : G \rightarrow D$ une fonction qui associe à tout objet de G sa description $\delta(g)$ de D. On appelle $(G, (D, \sqcap), \delta)$ une structure de patrons. Le inf-demi-treillis $(D, \sqcap)$ désigne l'*espace de recherche*. Lorsque l'ensemble d'objets G est décrit par un unique attribut de type intervalle numérique, on note par $V_[$ (resp. $V_]$) l'ensemble des bornes gauches (resp. droites) ordonné de manière croissante. on a : $\delta(g) = [g_[, g_]]$ pour $g \in G$, l'infimum entre $c, d \in D$ est donné par : $c \sqcap d = [c_[, d_]] \sqcap [c_[, d_]] = [min(c_[, d_[), max(c_], d_])]$. Nous nous intéressons à énumérer uniquement les motifs fermés, c'est à dire les descriptions $d \in D$ de support $supp(d) = S$ où $d^\square = S^\square$ avec $d^\square = \{g \in G \mid d \sqsubseteq \delta(g)\}$ et $S^\square = \bigsqcap_{g \in S} \delta(g)$ (en d'autres termes $d^{\square\square} = d$). Afin d'énumérer l'ensemble des motifs fermés fréquents (i.e. $|S| \geq minSup$), nous utilisons l'algorithme introduit par Kaytoue et al. (2011) qui parcourt l'ensemble des motifs fermés en partant du motif le plus général ($G^\square$) et en faisant un parcours sur *l'inf-demi-treillis* $(D, \sqcap)$ en profondeur. A chaque étape du parcours ($d = [a, b]$ est le motif de cette étape) : (i) on calcule sa fermeture $d^{\square\square}$ et on applique le *test de canonicité* qui permet de savoir si c'est bien la première fois que le motif fermé est généré. (ii) si le test réussi, on garde le motif fermé $d^{\square\square}$ puis on applique un changement minimal gauche ($CMG(d) = d_G = [suiv(a), b]$) et droite ($CMD(d) = d_D = [a, prec(b)]$) (après un CMD on a le droit de faire uniquement un CMD) avec $suiv(a)$ (resp. $pred(b)$) est l'élément qui vient juste après a dans $V_[$ (resp. vient juste avant b dans $V_]$) et on continue le parcours (avec un élagage éventuel selon $minSup$). Dans le cas où le test de canonicité échoue on procède à un retour arrière. *Le test de canonicité* permet de vérifier dans le cas où le changement minimal effectué sur d est un CMD ($d_D = [a, prec(b)]$) qu'après calcul de la fermeture ${d_D}^{\square\square} = [a_2, b_2]$ on a bien $a_2 = a$. C'est précisément ce test qui permet d'adapter la méthode de Kaytoue et al. (2011) pour les objets de type *intervalle numérique*.

Dans le cas général où les objets sont décrits par un ensemble d'attributs de type symbolique, numérique ou intervalle numérique, le parcours se déroule de la même manière comme montré précédemment à deux différences près : (i) la méthode de parcours des descriptions dépend du type de l'attribut (le cas des attributs symboliques (ou itemset) est traité par Kuznetsov (1993) ; et le cas des attributs numériques par Kaytoue et al. (2011)),(ii) un ordre canonique est établi sur les attributs, le test de canonicité vérifie sur la base de cet ordre qu'après fermeture d'une description obtenue à partir d'un changement minimal sur un attribut d'ordre i que toutes les restrictions sur les attributs $j < i$ n'ont pas changé. Le *test de canonicité* présenté auparavant s'applique uniquement dans le cas où l'attribut est de type intervalle numérique.

2 - Post-traitement sur les motifs contextuels extraits. La collection de motifs fermés extraite est très redondante. Plusieurs mécanismes d'élimination en post-traitement peuvent être utilisés pour réduire cette redondance. On utilise dans cet article **Top-k.**, une approche itérative qui consiste à ne retenir que les k-premiers motifs qui maximise une mesure de qualité (ou plusieurs dans un ordre donnée) tout en assurant une dissimilarité minimale. Nous utilisons pour cela une mesure de ressemblance : *Ressemblance*$(c, d) = \frac{|supp(c) \cap supp(d)|}{min(|supp(c)|, |supp(d)|)}$. Soit un seuil maximal de ressemblance toléré $seuil \in [0, 1]$, on procède ainsi tant que l'ensemble de top-k ne contient pas k éléments : on ajoute le motif de qualité maximale d au résultat final puis on supprime des candidats tout motif $c \neq d$ tel que *Ressemblance*$(c, d) \geq seuil$.

4 Expérimentations sur les traces de mobilité de véhicules

Nous disposons d'un ensemble de traces GPS issues de véhicules suite à une collaboration industrielle. On cherche à comprendre la mobilité de chaque individu *séparément*. Une trace est une séquence de triplets $(t, longitude, latitude)$ à partir de laquelle les points d'intérêts (*POI*) et les *visites* $< Poi, Jour\_de\_semaine, Intervalle\_de\_visite = [temps\_arrivee, temps\_depart] >$ sont extraits. Une visite qui s'étend sur plusieurs journées est découpée en plusieurs sous-visites.

Découverte de POI et extraction des visites. Tout d'abord, on nettoie et échantillonne les données afin de réduire le nombre d'événements. On cherche alors à pondérer les événements restants (ou les régions de l'espace), afin de donner plus d'importance aux points qui sont des points d'intérêt potentiels pour l'individu. La pondération consiste à calculer en chaque point de l'espace un poids en utilisant les événements qui sont à proximité en prenant en considération leur distribution temporelle. Une région dans l'espace dont les événements à proximité forment des visites continues dans le temps a un poids plus élevé comparé à celle qui a des événements isolés dans le temps. Une fois la pondération faite, les points de poids nul sont éliminés. L'ensemble des points d'intérêt est extrait depuis la nouvelle distribution en utilisant une extension de l'algorithme *DBSCAN* (Ester et al. (1996)) qui prend en compte les poids calculés de chaque point en remplaçant la notion de *minPts* par *minWeight*. Ensuite, la base des visites est construite. Nous appliquons notre approche sur une trace de mobilité d'une durée de 7 semaines contenant *223 386 événements*. 3 POI et 1 POI infréquent (d'identifiant *I* qui contient les POI détectés mais visités une seule fois) ressortent. Au final 179 visites de durée $5mn$ sont construites en 20 secondes (processeur 2.60 GHZ et 4Go RAM, utilisé pour tous résultats rapportés par la suite).

Extraction des motifs. Nous utilisons comme mesure de qualité la *mesure de sédentarité* (à minimiser). Nous utiliserons un *support minimum* de *10* et les bornes des intervalles de visite sont arrondis à la dizaine de minute la plus proche. 22589 motifs fermés fréquents sont extraits en $11s$. Le top-3 des motifs de sédentarité sont donnés dans la Table 2 : on observe une redondance des descriptions. Après avoir appliqué la méthode itérative de réduction de redondance avec une seuil de ressemblance maximale toléré égale à 30%, le deuxième et le troisième motif seront retirés du résultat et les trois meilleurs motifs sont indiqués par la Table 2 (bas). Le premier motif est de pureté maximale. En d'autres termes, les journées de dimanche à vendredi dans l'intervalle de temps [00:00:00, 10:39:59], Il y a de fortes chances que l'individu soit dans le POI 1 (probablement l'habitat de l'individu). On constate que le deuxième meilleur motif est pur pour le POI 2. Les journées et les horaires concernées corresponds à des horaires de travail. La sémantique du POI 2 peut être ainsi le lieu de travail.

Motif	Poi visités	Entropie	Support
<['D', 'J', 'L', 'M', 'Me', 'V'],[00:00:00, 10:39:59]>	1	0	37
<['D', 'J', 'L', 'M', 'Me', 'V'],[00:00:00, 10:29:59]>	1	0	36
<['J', 'L', 'M', 'Me', 'V'],[00:00:00, 10:29:59]>	1	0	35

Motif	Poi visités	Entropie	Support
<['D', 'J', 'L', 'M', 'Me', 'V'],[00:00:00, 10:39:59]>	1	0	37
<['L', 'M', 'V'],[09:50:00, 19:49:59]>	2	0	12
<['J', 'M', 'Me'],[09:50:00, 19:09:59]>	2,I	0,414	11

TAB. 2 – *Top-3 des motifs (haut) et Top-3 avec un seuil de ressemblance* 30% *(bas)*

5 Conclusion

Nous nous sommes intéressés à la compréhension d'une trace de mobilité à travers un ensemble de motifs locaux qui la couvrent. Pour cela, la trace est fragmentée en une base de visites de points d'intérêt, chaque visite est décrite par un contexte, et nous cherchons des motifs contextualisés pertinents. Nous avons étendu un algorithme existant pour gérer le cas des attributs de type intervalle. Enfin, des expérimentations préliminaires montrent l'intérêt de notre approche.

Références

Ester, M., H. Kriegel, J. Sander, et X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pp. 226–231. AAAI Press.

Gambs, S., M. Killijian, et M. N. del Prado Cortez (2011). Show me how you move and I will tell you who you are. *Transactions on Data Privacy 4*(2), 103–126.

Ganter, B. et S. O. Kuznetsov (2001). Pattern structures and their projections. In *Conceptual Structures : Broadening the Base*, pp. 129–142. Springer.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis*. Berlin : Springer.

Gonzalez, M. C., C. A. Hidalgo, et A.-L. Barabasi (2008). Understanding individual human mobility patterns. *Nature 453*(7196), 779–782.

Kaytoue, M., S. O. Kuznetsov, et A. Napoli (2011). Revisiting numerical pattern mining with formal concept analysis. In *Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pp. 1342–1347.

Kuznetsov, S. (1993). A Fast Algorithm for Computing all Intersections of Objects in a Finite Semi-lattice. *Automatic Documentation and Mathematical Linguistics 27*(5), 11–21.

Novak, P. K., N. Lavrac, et G. I. Webb (2009). Supervised descriptive rule discovery : A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research 10*, 377–403.

Trasarti, R., F. Pinelli, M. Nanni, et F. Giannotti (2011). Mining mobility user profiles for car pooling. In *Int. Conf. on Knowledge discovery and data mining*, pp. 1190–1198. ACM.

Zhang, J.-D., J. Xu, et S. S. Liao (2013). Aggregating and sampling methods for processing gps data streams for traffic state estimation. *IEEE Transactions on Intelligent Transportation Systems 14*(4), 1629–1641.

Summary

Whereas several works propose to characterize a mobile object in a global way, few researchers have addressed the problem in a local way. Given the mobility trace of a mobile object, the method we introduce allows to discover contextualized mobility patterns: each pattern corresponds to a mobility context (e.g. *a sunny Friday*) and is provided with some measures with a semantics (the pattern describes sedentary or exceptional phase). Experiments on mobility traces show promising results.