Recommandation de séquences d'activités lors d'évènements distribués

Diana Nurbakova, Léa Laporte, Sylvie Calabretto, Jérôme Gensel

Recommandation de séquences d'activités lors d'évènements distribués

Diana Nurbakova[*], Léa Laporte[*], Sylvie Calabretto[*], Jérôme Gensel [**]

[*] *LIRIS UMR 5205, INSA Lyon, France*
[**] *LIG UMR 5217, Université Grenobles Alpes, France*

RÉSUMÉ. *Le nombre d'événements sociaux augmente de manière significative et les services basés sur la localisation deviennent partie intégrante de notre vie. Ainsi la recommandation de séquences d'activities devient une application émergente importante. Ce problème est crucial dans le cas d'événements distribués (e.g. festival ou croisière) qui rassemblent plusieurs activités concurrentes. Un participant à de tels événements est submergé par le choix de nombreuses activités possibles et fait face au problème de sélection d'activités. Dans cet article, nous formulons le problème de recommandation de séquences d'activités comme la combinaison du problème de recommandation d'événements et du problème de planification. Nous proposons une nouvelle approche qui évalue l'intérêt de l'utilisateur pour une activité fondée sur diverses facettes, explore l'historique de l'utilisateur pour extraire leurs motifs comportementaux et les utilise dans la construction de l'itinéraire. Une évaluation sur un jeu de données construit à partir du programme d'une croisière montre une amélioration moyenne de 9.7% par rapport à l'état de l'art. Ce travail a été présenté à la conférence ICCS 2017 (Nurbakova* et al.*, 2017).*

ABSTRACT. *As the amount of social events significantly increases and location-based services become an integral part of our life, being able to recommend sequences of activities emerges as an important application. This problem is crucial in the case of distributed events (e.g. festival or cruise) that gather multiple competitive activities. An attendee of such events is overwhelmed with numerous possible activities and faces the problem of activity selection. In this paper, we formulate the problem of activity sequences recommendation as a combination of event recommendation and scheduling problem. We propose a novel approach that evaluates the user's interest in an activity based on various facets, mines the users' historical traces to extract their behavioural patterns and uses them in the construction of the itinerary. Evaluation on a dataset built over a cruise program shows an average improvement of 9.7% over the state-of-the-art.*

MOTS-CLÉS : *Activités spatio-temporelles, recommendation de séquences d'activités, incertitudes*

KEYWORDS: *Spatio-temporal activities, recommendation of activity sequences, uncertainty*

1. Introduction

Notre travail s'intéresse au problème émergent de la recommandation de séquences ou planning d'activités lors de grands événements distribués, tels des croisières, festivals et conventions, constitués d'un grand nombre de sous-événements ou activités de courte durée se déroulant en parallèle. Ce type d'événements attire chaque année des millions de participants et ce chiffre va en augmentant. Ainsi, selon le rapport annuel du CLIA en 2017[1], association regroupant les acteurs du domaine des croisières maritimes et fluviales, le nombre global de participants à des croisières a augmenté de 30% entre 2009 et 2015, pour atteindre plus de 23 millions de voyageurs en 2015. Les croisières constituent un bon exemple d'événement distribué, les organisateurs de croisière proposant généralement un large panel d'activités à leurs usagers, qui font alors face à un problème de sélection des activités. Considérons l'exemple suivant : Scoby profitent de ces vacances à bord d'une croisière de 7 nuits dans les Caraïbes. Chaque jour, il doit faire son choix parmi une centaine d'activités différentes d'une durée moyenne de 45 minutes. A chaque instant, environ cinq activités ont lieu en parallèle. Quel est la meilleure façon pour Scoby de planifier sa journée afin d'en profiter le plus possible ?

Le processus de prise de décision associé au choix des activités est un problème complexe, en particulier dans le cas de grands événements distribués. Par conséquent, le problème de recommandation de planning d'activités est lui aussi complexe et doit tenir compte d'un grand nombre de sources d'incertitudes. Nous décrivons les principales sources et types d'incertitudes, regroupées selon la classification de (Tchernykh *et al.*, 2015), dans le tableau 1.

Le problème de la recommandation des séquences d'activités survient dans divers domaines comme la planification des conférences, des festivals (*e.g.* ComicCon), et de grands événements distribués (Schaller *et al.*, 2013), la planification de séjours touristiques ou de vacances (Gavalas *et al.*, 2014), et le crowdsourcing mobile (Fonteles *et al.*, 2015). Dans cet article, une croisière est considérée comme scenario d'application. Nous considérons les environnements intérieurs et les environnements extérieurs restraint, ce qui implique que le temps de trajet est moins important que dans le cas d'une recommandation de circuits touristiques et de Points d'Intérêt (POI).

Le principal enjeu de la recommandation des séquences d'activités repose sur le fait que les activités sont uniques, à durée restreinte et ont lieu dans le futur. Ainsi, chaque fois qu'un utilisateur souhaite choisir une activité à effectuer, il / elle doit détecter une activité qu'il / elle peut préférer parmi toutes les alternatives se produisant pendant le même intervalle de temps, en tenant compte du fait qu'il / elle ne peut pas rejoindre une autre activité d'intérêt en raison des contraintes de disponibilité / temps. Par conséquent, l'objectif n'est pas seulement de définir l'intérêt des utilisateurs pour les événements à venir, mais de fournir un programme personnel (itinéraire) par jour

1. https ://www.cruising.org/docs/default-source/research/clia-2017-state-of-the-industry.pdf ?sfvrsn=6

Sources d'incertitudes	Type d'incertitude	Description
Données	Retour implicite	Nous ne disposons d'aucun retour explicite de pertinence sur les données collectées, en particulier, aucune note ou évaluation n'est disponible pour les activités
	Occurrence future	Les activités à recommander ont par nature une durée de vie limitée et ont lieu dans le futur, limitant les informations disponibles : aucun retour d'expérience n'est disponible au moment de la recommandation, qui doit être effectuée avant que l'activité ne débute
	Contraintes	Les items de recommendation sont de courte durée et sont disponibles à un moment précis et en un lieu précis, ce qui fait qu'un utilisateur préfère une activité à une autre pour une plage de temps donnée
	Biais de participation	Un utilisateur peut rejoindre une activité qui ne l'intéresse pas, et toutes les activités qui manquent à un utilisateur ne représentent pas un intérêt personnel pour lui
Hypothèses	Maximisation de la satisfaction	La satisfaction de l'utilisateur est souvent considérée comme étant cumulative, *i.e.* plus vous rejoignez d'activités, plus vous êtes satisfait
Paramètres	Estimation des paramètres	La nature probabiliste des paramètres du modèle a un impact sur la qualité de la solution
Modèle	Profilage	Le type de représentation des caractéristiques des alternatives et des préférences des utilisateur, ainsi que la manière de les maintenir à jour, ont un impact sur la compréhension des processus en cours de prise de décision des utilisateurs
	Score de préférence	La qualité d'estimation des préférences dépend de choix des caractéristiques considérées
	Construction d'itinéraires	La fonction objectif choisi a une influence sur la planification des séquences d'activité

Tableau 1. *Sources d'incertitudes pour la recommandation de séquences d'activités*

des activités. Les activités étant uniques et à venir, il est nécessaire de pouvoir les recommander sans retour explicite des utilisateurs et sans aucune information externe (*e.g.* avis ou commentaires des utilisateurs). Par ailleurs, il s'agit de traiter des préférences utilisateurs très incertaines et de s'assurer qu'un utilisateur sera en mesure d'assister à toutes les activités sélectionnées à temps. Nous supposons pour ce faire que nous avons accès aux lieux et points d'intérêts précédemment visités par les utilisateurs, ainsi qu'à la date de visite.

Dans cet article, nous nous intéressons à la question de recherche suivante : *Comment prédire et maximiser la satisfaction des utilisateurs relative à une séquence d'activités proposée, étant donnée leur expérience passée ?* Nous décomposons cette question en trois sous-problèmes :

Q1 : Comment extraire les préférences des utilisateurs à partir de données incertaines et comment évaluer l'intérêt des utilisateurs pour une activité à venir, étant donné le peu d'informations à son sujet ?

Q2 : Comment récupérer les motifs comportementaux des utilisateurs à partir des données historiques, afin de proposer une recommandation adaptée à leurs habitudes et préférences ?

Q3 : Comment organiser les activités en une séquence qui maximise la satisfaction de l'utilisateur tout en tenant compte des contraintes spatio-temporelles et de la nature séquentielle des activités ?

La présente étude est liée aux domaines de recherche suivants : la recommandation de Points d'intérêt (POI), la recommandation d'événements, la recommandation de voyages ou circuits touristiques et la construction d'itinéraires ou planning. *La recommandation de POI* vise à proposer aux utilisateurs des listes de k points d'intérêt tenant compte de leurs préférences, la composante géographique étant considérée comme la plus importante (Li *et al.*, 2015), bien que certains travaux exploitent aussi les influences catégorielles (type ou thème de l'activité) ou les influences sociales (Zhang et Chow, 2015a). La principale limite des techniques de recommandation de POI est qu'elles ne tiennent pas compte des contraintes de disponibilité des POI, du temps de trajet entre les POI, de la durée des visites, du budget temporel de l'utilisateur ni de l'ordre de visite. Des travaux récents se sont intéressés à la *recommandation d'évènements*, en particulier dans les réseaux sociaux basés sur les évènements comme Meetup. Ainsi, (Macedo *et al.*, 2015) propose d'exploiter, en plus des signaux textuels et collaboratifs, des signaux complémentaires tels que le réseau social ou les préférences géographiques et temporelles des utilisateurs, qu'ils combinent au sein d'une approche d'apprentissage d'ordonnancement afin de construire une liste ordonnée d'évènements. À l'instar des techniques de recommandation de POI, cette méthode ne prend pas en compte la disponibilité limitée des événements et l'ordre relatif de déroulement des événements. *La recommandation de séjours et circuits touristiques* vise à fournir à un utilisateur une séquence de POI à visiter, en tenant compte des contraintes spatiales et temporelles. La plupart des études récentes (Zhang et Chow, 2015b) décompose le problème en deux parties : (1) une estimation des scores individuels pour chaque POI, (2) la construction d'un itinéraire réalisable composé des POI les plus adaptés (au sens du score d'intérêt) pour un utilisateur donné (Gavalas *et al.*, 2014). Contrairement au problème traité dans ce document, le problème de recommandation de circuits touristiques ne tient pas compte du caractère unique et éphémère des activités. Enfin, le problème de la *construction d'itinéraires* est souvent modélisé comme une instance du Problème d'Orientation (OP) ou d'une de ses variantes (Vansteenwegen *et al.*, 2011). OP vise à déterminer un chemin hamiltonien limité par un certain budget temporel maximisant le score cumulé en visitant

Activité : *The Comedy & Hypnosis of Ricky Kalmon*
Lieu : Walt Disney Theatre, $l = (0, 880, 0)$; **Time window** : Day 3, 23 :00-23 :45, $t = (1435014000, 1435016700)$; **Duration** : $\delta = 2700$; **Catégories** : Adults, Variety Show
Description : Featuring the Comedy & Hypnosis of Ricky Kalmon, as he entertains you in this adult exclusive show.

Tableau 2. *Un exemple d'activité*

chacun des sommets. Ces approches ne s'intéressent pas à la façon dont les scores des sommets ont été calculés et se focalisent sur la construction de l'itinéraire.

Dans cet article, nous proposons une approche originale pour résoudre le problème de recommandation de séquences d'activités. Les contributions clés de notre travail sont : une définition formelle du problème de recommandation de séquences d'activité ; une méthode intégrale pour calculer les scores d'activité ; un algorithme pour extraire les motifs comportementaux des utilisateurs à partir des activités passées ; une intégration des motifs de transition typiques des utilisateurs sur la construction de l'itinéraire. Par ailleurs, en l'absence de jeu de données de référence, nous avons mené une étude utilisateurs afin de créer un jeu de données d'évaluation adapté à la tâche.

Cet article est organisé comme suit. La section 2 définit le problème et introduit les notations. Dans la section 3, nous présentons notre approche qui consiste en trois parties : le calcul des scores d'activité, l'extraction de motifs comportementaux, et la construction de l'itinéraire, ou séquence d'activités. La section 4 décrit le protocole expérimental, y compris le jeu de données utilisé pour l'évaluation. La section 5 présente les résultats obtenus. Enfin, la section 6 conclut l'article et présente les perspectives de nos travaux. Cette soumission est une traduction en français d'un article publié dans la conférence internationale "International Conference on Computational Science" (ICCS) 2017 (Nurbakova *et al.*, 2017).

2. Formulation du problème

Dans cette section, nous définissons les notations utilisées dans l'article et nous formulons le problème.

Une *activité* $a = \langle l, t, \delta, c, d \rangle$ est un événement auquel un utilisateur u peut assister dans un certain lieu et sur une certaine plage horaire. Elle se caractérise par sa localisation (lattitude, longitude, altitude), $l = (x, y, z)$, sa fenêtre de temps ou plage horaire (heure de début t_s et heure de fin t_e) représentant sa disponibilité $t = (t_s, t_e)$, sa durée δ, un vecteur de catégories associé $c = (c_1, ..., c_k)$, et une description d. Un exemple d'activité est donné dans le tableau 2. $A = \{a_1, a_2, ..., a_N\}$ représente l'ensemble des activités disponibles.

Une *séquence d'activité* (ou *itinéraire*) $\xi^u = \left(a^u_{(s)}, ..., a^u_{(s+k)}\right)$, où $1 \leq s \leq s+k \leq N$, est une série ordonnée d'activités pour un utilisateur u, mettant en jeu un ensemble de contraintes spatio-temporelles [2] telles que la contrainte de disponibilité et la contrainte de budget temporelle (l'utilisateur a-t-il suffisamment de temps libre restant pour réaliser la ou les activités ?). *La contrainte de disponibilité d'une activité* spécifie qu'une activité a_i peut être effectuée seulement pendant sa période de disponibilité, limitée par sa date de début t_s et sa date de fin t_e, *i.e.* $t_s \leq start(a_i) \leq t_e$. Ici, $start(a_i)$ représente la date à laquelle un utilisateur commence à assister à l'activité a_i, sachant qu'il/elle peut rejoindre l'activité a_i lorqu'elle devient disponible et une fois qu'il / elle a fini d'effectuer l'activité précédente et s'est déplacé vers le lieu de l'activité courante, *i.e.* $start(a_{(i)}) = \max\{start(a_{(i-1)}) + \delta(a_{(i-1)}) + time(a_{(i-1)}, a_{(i)}), t_s(a_{(i)})\}$, où $time(a_{(i-1)}, a_{(i)})$ est le temps de déplacement pour aller du lieu d'activité $a_{(i-1)}$ au lieu d'activité $a_{(i)}$. *La contrainte de budget temporel* limite le temps total nécessaire pour suivre toutes les activités d'un itinéraire, en incluant la durée des activités et le temps de déplacement avec le budget temporel maximal T_{max} donné, correspondant au temps pendant lequel l'utilisateur est disponible pour effectuer ses activités. Il peut être défini par un utilisateur.

Une *fonction de satisfaction* $r\left(a_i, u\right),\ r : A \longrightarrow \Re^+$ caractérise l'appariement entre l'activité a_i avec l'intérêt d'un utilisateur u. La satisfaction avec un itinéraire ξ^u pour un utilisateur u est définie comme la somme des scores des activités de l'itinéraire, $r\left(\xi^u, u\right) = \sum_{a_i \in \xi^u} r\left(a_i, u\right)$.

Le problème de recommandation de séquences d'activités consiste à trouver un itinéraire ξ^u qui maximise la satisfaction de l'utilisateur $r\left(\xi^u, u\right)$, étant donné un utilisateur u et un ensemble d'activités $A = \{a_i\}_{i=\overline{1,N}}$.

Dans cet article, nous effectuons les *hypothèses* suivantes : (1) *Satisfaction maximale* : dans le contexe de la présence des utilisateurs à un événement distribué, leur objectif est d'obtenir le maximum de satisfaction de l'expérience globale. (2) *Traçabilité des utilisateurs* : l'historique des expériences passées des utilisateurs, constitué d'ensembles de coordonnées géospatiales et d'horodatages, est disponible. (3) *Déplacement dans l'espace* : le temps de déplacement des utilisateurs entre les lieux des activités est fonction de la distance. Nous supposons que tous les utilisateurs se déplacent à vitesse constante.

3. Proposition

Nous proposons une approche intégrale pour la recommandation de séquences d'activités exploitant les intérêts des utilisateurs, l'influence séquentielle, les contraintes spatiales et temporelles. Elle se compose de trois parties principales : 1.

2. Ici, nous utilisons des parenthèses en indice pour indiquer que les éléments à l'intérieur d'une séquence sont ordonnés, de telle sorte qu'ils peuvent être différentiés des éléments de l'ensemble complet.

Le calcul des scores personnalisés pour chaque activité, 2. L'extraction des motifs de comportement de l'utilisateur, 3. La construction de l'itinéraire en utilisant les données fournies par les étapes précédentes.

3.1. *Calcul des scores personnalisés*

Nous étudions maintenant notre **Q1 :** *Comment extraire les préférences des utilisateurs à partir de données incertaines et comment évaluer l'intérêt des utilisateurs pour une activité à venir, étant donné le peu d'informations à son sujet ?*. Les activités sont uniques et ne peuvent être effectuées que pendant leur fenêtre temporelle. Il n'y a pas d'évaluations ou de notes des activités à venir qui pourraient être utilisées pour prédire les scores. Des traces d'utilisateurs sont disponibles (voir l'hypothèse de *Traçabilité des utilisateurs*) et peuvent, en combinaison avec le programme d'activités, être utilisées pour déterminer les activités auxquelles un utilisateur a participé dans le passé. En raison de l'absence de commentaires explicites sur le degré d'intérêt de l'utilisateur pour une activité suivie, la présence de l'utilisateur à une activité est considérée comme un retour implicite positif d'intérêt. Dans ce travail, nous proposons d'explorer le contenu (description) des activités ainsi que les influences catégorielles et temporelles afin d'estimer l'intérêt des utilisateurs envers les activités.

Influence du contenu : Nous considérons l'influence textuelle en appliquant un modèle de sac de mots à la description des activités. Chaque activité a est représentée par son vecteur TF-IDF noté $\vec{e}$ afin de construire des profils utilisateurs positifs et négatifs. Le profil positif d'un utilisateur U, noté U_{pos}, consiste en un résumé des vecteurs TF-IDF des descriptions des activités effectuées par l'utilisateur, tandis que le profil négatif de l'utilisateur U, noté U_{neg} est construit sur des activités non effectuées. Le score d'une activité future est calculé comme une combinaison linéaire des mesures de similarité cosinus entre le vecteur de l'activité et les profils positif et négatif, de telle sorte que la similarité avec des activités non effectuées soit utilisée comme une pénalité. Notre hypothèse est que les activités non effectuées présentent l'intérêt moins certain pour l'utilisateur. Le score est ainsi calculé de la façon suivante : $\hat{r}_{cb}(a,u) = \alpha_u \cdot cos(U_{pos}, \vec{e}) - \beta_u \cdot cos(U_{neg}, \vec{e})$. Les paramètres $alpha_u$ et $beta_u$ sont choisis pour un utilisateur donné en optimisant la fonction de perte par validation croisée 10-fold.

Influence Catégorielle : Nous proposons d'utiliser les catégories d'activités déjà effectuées par l'utilisateur afin d'estimer l'intérêt d'un utilisateur pour une activité future d'une certaine catégorie. Chaque activité a est associé à une liste de catégories $\mathcal{C}_a = \{c_j\}$. Ainsi, pour chaque utilisateur et chaque catégorie, nous calculons la fréquence d'une catégorie parmi les activités passées de l'utilisateur : $freq(c_i, u) = \frac{|A_{u,c_i}| \cdot w_a}{|A_u|}$, où $|A_{u,c_i}|$ est le nombre d'activités effectuées par un utilisateur u qui appartiennent à la catégorie c_i, $w_a = 1/|\mathcal{C}_a|$ est un poids lié au nombre de catégories à laquelle une activité $a \in A_{u,c_i}$ est associée, et $|A_u|$ est le nombre total d'activités effectuées par l'utilisateur u. Étant donné un utilisateur et une activité, nous représentons le score catégoriel d'une activité comme la somme des fréquences catégorielles correspondantes, *i.e.* $\hat{r}_{cat}(a,u) = \sum_{\mathcal{C}_a} freq(c_j, u)$.

Stratégie 1 : All-at-once ;	Stratégie 2 : Day-after-Day ;
Input : User's Attendance Matrix $\mathcal{M}$, New activities $NewEvent$;	**Input :** User's Attendance Matrix $\mathcal{M}$, New activities $NewEvent$, Number of past days $PastDays$, Total number of days $DayNum$;
Output : Activities scores $\mathcal{R}$;	**Output :** Activities scores $\mathcal{R}$;
Calculate $\mathcal{R}(NewEvents, \mathcal{M})$	**Initialisation** $\mathcal{M}^{(0)} \leftarrow \mathcal{M}$;
	for $i \leftarrow PastDays$ **to** $DayNum$ **do**
	Calculate $\mathcal{R}^{(i)}(NewEvent^{(i)}, \mathcal{M}^{(i)})$;
	$\mathcal{M}^{(i)} \leftarrow \mathcal{M}^{(i)} \cup \mathcal{R}^{(i)}$;
	$i \leftarrow i + 1$
	end

Tableau 3. *Stratégies 1 et 2 pour estimer les scores d'intérêt des activités*

Influence temporelle : Un autre facteur pouvant influencer les utilisateurs dans leur décision de rejoindre une activité est l'aspect temporel, *i.e.* le moment où l'activité a lieu. Notre intuition est qu'un utilisateur est plus actif ou plus disponible pour effectuer des activités à certaines périodes de la journée qu'à d'autres. Pour formaliser cette intuition, nous divisons un jour en intervalles de temps de 15 minutes, soit 96 intervalles par jour. Nous représentons ensuite chaque activité comme un vecteur binaire t_a de dimension 1×96 tel qu'un élément égal à 1 indique que la fenêtre de temps de disponibilité d'une activité inclut cet intervalle de temps. Un utilisateur est alors représenté comme le vecteur binaire construit sur l'union des intervalles de temps de ses activités passées t_u . Le score temporel, adapté des travaux de (Sang *et al.*, 2015), est défini sur la base des relations temporelles entre un vecteur de disponibilité d'une activité future et le profil temporel d'un utilisateur :

$$\hat{r}_{time}(a,u) = \begin{cases} 1, & \text{if } t_a \cap t_u \\ 0.5, & \text{if } t_a \cap \{t_u - 1 \cup t_u + 1\} \\ 0.1, & \text{otherwise} \end{cases}.$$

Combinaison des influences : Pour améliorer l'efficacité de la recommandation, nous proposons de combiner les trois influences précédentes, de deux façons différentes. Dans une première approche, nous proposons de définir *un score hybride* (LinC) : $\hat{r}_{hyb}(u,a) = (\gamma_u \cdot \hat{r}_{cb}(u,a) + \delta_u \cdot \hat{r}_{cat}(a,u)) \cdot \hat{r}_{time}(a,u)$, où γ_u et δ_u sont fixés en optimisant la fonction de perte par validation croisée 10-folds. Dans une deuxième approche, nous proposons d'utiliser un algorithme de régression logistique pour combiner les scores catégoriels, textuels et temporels. La probabilité d'affecter une activité à la classe 1 constitue notre *score de régression logistique* (LogR).

Stratégies d'estimation des scores : Nous proposons d'organiser le processus d'estimation de scores selon deux stratégies que nous appelons *Stratégie 1* ('All-at-Once') et *Stratégie 2* ('Day-after-Day'). La table 3 représente les pseudocodes des algorithmes. Ainsi, la Stratégie 1 considère toutes les activités futures simultanément et

estime leurs scores par rapport aux modèles précédemment présentés. La Stratégie 2 estime les scores des activités sur une base quotidienne. L'expérience utilisateur, pour un jour donné, est enrichie à l'aide des scores estimés pour le jour précédent. Ces données enrichies sont utilisées dans la prochaine itération de l'algorithme comme données de l'historique.

3.2. *Extraction des motifs de comportement des utilisateurs*

Cette section s'intéresse à notre question **Q2 :** *Comment récupérer les motifs comportementaux des utilisateurs à partir des données historiques, afin de proposer une recommandation adaptée à leurs habitudes et préférences ?*. Notre objectif est de récupérer les transitions les plus typiques entre les activités consécutives, *i.e.* les sous-séquences des activités des utilisateurs ou motifs de comportement des utilisateurs. Deux activités forment une séquence si l'intervalle de temps entre la fin de première et le début de la seconde est inférieur au seuil fixé (Zhang *et al.*, 2014).

Nous proposons de construire un *graphe de transition activité-activité* (A^2TG) et un *graphe de transition catégorie-catégorie* (C^2TG) en étendant le concept de graphe de transition lieu-lieu proposé par (Zhang *et al.*, 2014) pour modéliser les transitions entre POIs. Le graphe A^2TG modélise les transitions entre activités représentées par leur date et leur lieu. Comme les activités sont uniques, le graphe A^2TG ne peut pas être utilisé directement pour l'estimation des probabilités de transition entre nouvelles activités. Pour contourner ce problème, nous les utilisons pour construire le graphe C^2TG qui modélise les transitions généralisées entre catégories, et qui servira de base pour estimer les probabilités de transitions entre nouvelles activités.

Les séquences d'activités extraites des traces de l'utilisateur et du programme d'un événement distribué sont utilisées pour construire le graphe A^2TG. Les noeuds du graphe, $V = \{a_1, ..., a_N\}$, correspondent aux activités effectuées par un utilisateur. A chaque noeud du graphe nous attribuons une valeur $InCount(a_i)$ qui représente le nombre d'arêtes entrantes. Notre intuition est que la satisfaction de l'utilisateur dépend plus de ses expériences passées que de ses expériences futures. Ceci différencie notre approche de celle de (Zhang *et al.*, 2014) qui attribue le nombre d'arêtes sortantes. Les arêtes représentent les transitions entre les activités et sont associées au nombre de transitions, $TransCount(a_i \rightarrow a_j)$. Nous proposons ensuite de passer au niveau de la catégorie en supposant que les catégories d'activités sont connues.

C^2TG est construit de la même manière. Ses noeuds représentent les catégories c_i associées aux activités effectuées et sont caractérisés par le nombre d'arêtes entrantes, $InCount(c_i)$, calculé comme suit : $InCount(c_i) = \sum_{a_j \in c_i} InCount(a_j)$. Les arêtes représentent des transitions entre les catégories et sont associées au nombre de transitions, $TransCount(c_i \rightarrow c_j)$, qui est calculé en utilisant le $TransCount$ des activités correspondantes comme suit : $TransCount(c_i \rightarrow c_j) = \sum_{\substack{a_k \in c_j, \\ a_g \in c_i}} TransCount(a_k \rightarrow a_g)$. Etant donné C^2TG, nous esti-

Séquence d'activités : {*Crafts : Door Hangers* → *Pictionary Challenge* → *Singles' Lunch* → *Goofy* → *The Comedy and Hypnosis of Ricky Kalmon}*
Séquence de catégories : {*Fun for All Ages* (F) → *Fun for All Ages* (F) → *Adults* (A) → *Characters* (C) → *Fun for All Ages* (F)}

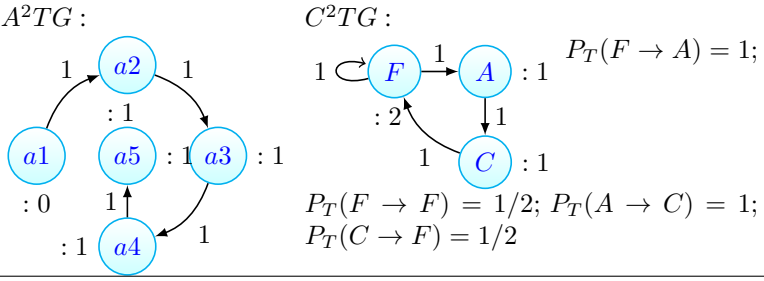

Tableau 4. *Exemple de A^2TG, C^2TG et P_T. Les labels de noeud représentent les valeurs de $InCount$.*

mons la probabilité de transition d'une catégorie à une autre comme suit :

$$P_T(c_i \to c_j) = \begin{cases} \frac{TransCount(c_i \to c_j)}{InCount(c_j)}, & \text{if } InCount(c_j) \neq 0 \\ 0, & \text{if } InCount(c_j) = 0 \\ & \quad and\ c_i \neq c_j \\ 1, & \text{if } InCount(c_j) = 0 \\ & \quad and\ c_i = c_j \end{cases}$$

La probabilité de transition entre deux activités est estimée selon le processus inverse. Un exemple illustratif est donné dans le tableau 4.

3.3. *Construction de la séquence d'activités*

Dans cette section, nous nous focalisons sur la question **Q3 :** *Comment organiser les activités en une séquence qui maximise la satisfaction de l'utilisateur tout en tenant compte des contraintes spatio-temporelles et de la nature séquentielle des activités ?* Etant donné un utilisateur, un ensemble d'activités A définies par leurs localisations données par leurs coordonnées, la fenêtre de temps de leurs disponibilité, leur durée, les scores d'intérêt personnalisés, le temps de déplacement entre deux localisations, nous cherchons une séquence d'activités qui maximise le score total collecté, *i.e.* la satisfaction de l'utilisateur à partir des activités déjà effectuées. La construction d'itinéraires peut être formulée comme un problème d'orientation avec des fenêtres de temps (Orienteering Problem with Time Windows - OPTW) (Vansteenwegen *et al.*, 2009).

Vansteenwegen *et al.* (Vansteenwegen *et al.*, 2009) ont proposé un algorithme itératif de recherche locale, Iterated Local Search (ILS), pour résoudre l'OPTW. ILS est un algorithme heuristique qui, à chaque itération, cherche un noeud à ajouter à la solution actuelle qui va maximiser le score total de l'itinéraire. Une vérification

de la validité de l'ajout d'un noeud est effectuée pour garantir que l'insertion d'un nouveau noeud respecte les contraintes temporelles. Ainsi, à chaque itération, pour chaque noeud candidat a_k la valeur suivante est calculée : $Ratio_k = \frac{\hat{r}_k^2}{Shift_k}$, où $\hat{r}_k$ est le score du noeud et $Shift_k$ est un décalage temporaire, produit par l'insertion de l'activité a_k dans le parcours actuel entre les noeuds a_i and a_j, c'est-à-dire : $Shift_k = t_{ik} + Wait_k + \delta_k + t_{kj} - t_{ij}$, où $t_{..}$ est le temps de déplacement entre deux noeuds, $Wait_k$ est le temps d'attente du début de l'activité à l'arrivée au noeud a_k, δ_k est la durée de l'activité a_k, i et j sont les indices des noeuds entre lesquels le noeud candidat a_k sera ajouté. Ainsi ILS ajoute un noeud valide a_k avec le ratio $Ratio_k$ maximal. Pour la formulation mathématique du problème OPTW et plus de détails sur ILS voir (Vansteenwegen *et al.*, 2009).

Nous proposons une adaptation de l'algorithme ILS. Ainsi, nous proposons d'ajuster la valeur de $Ratio_k$ avec la probabilité de transition de l'activité précédente a_{k-1} à l'activité actuelle a_k comme suit : $Ratio_k = \frac{\hat{r}_k * P_T(a_{k-1} \to a_k)}{Shift_k}$, où $\hat{r}_k$ est le score de l'activité a_k, $P_T(a_{k-1} \to a_k)$ est la probabilité de transition entre a_{k-1} et a_k. Nous appelons notre approche ILS_TP. Notre intuition est que l'intégration de la probabilité de transition, c'est-à-dire des motifs séquentiels de comportement d'utilisateurs, permettra d'améliorer la construction de l'itinéraire.

4. Protocole expérimental

4.1. *Construction du jeu de données*

A notre connaissance, il n'existe pas de jeu de données disponible pour la recommandation d'itinéraires lors d'évènements distribués. Nous avons donc créé un jeu de données sur lequel nous avons évalué nos approches en mode hors ligne. Ce jeu de données simule la participation d'utilisateurs à une croisière et a été créé de la façon suivante. Nous avons collecté les plannings d'activités proposés lors d'une croisière Disney de 7 nuits sur la mer des Caraïbes[3]. A chaque activité est généralement associée une description détaillée. Lorsque cette dernière était absente, en particulier pour les activités faisant référence à un film ou un personnage Disney, nous avons enrichi les descriptions à l'aide de données extraites d'IMDb (http ://www.imdb.com/) et wikia (http ://disney.wikia.com/) respectivement. Nous avons ensuite créé une enquête en ligne comportant trois parties afin d'acquérir des informations sur les utilisateurs. La première partie consiste en l'acquisition du profil personnel d'un utilisateur, et vise à collecter des informations comme le genre, l'expérience en matière de croisière, le type de groupe (famille, ami, etc.) avec lequel voyage l'utilisateur. La deuxième partie du questionnaire permet de collecter l'intérêt de l'utilisateur relatif aux activités proposées. Etant donnée une liste des activités proposées lors de la croisière, mais sans

3. http ://disneycruiselineblog.com/2015/07/personal-navigators-7-night-eastern-caribbean-cruise-on-disney-fantasy-itinerary-a-june-20-2015/

Statistique	# Activités	# Jours	# Utilisateurs	# Lieux	# Catégories
Valeur	595	7	23	47	52

Tableau 5. *Caractéristiques du jeu de données*

information horaire, extraites des plannings collectés précédemment, les répondants à l'enquête devaient indiquer leur intérêt pour les différentes activités en donnant une note sur une échelle de 0 (ne souhaite pas participer) à 5 (souhaite absolument participer). La troisième et dernière partie du questionnaire consistait à collecter les plannings d'activités désirés par les utilisateurs pour chaque jour de croisière. Etant donné une liste d'activités avec leurs créneaux horaires, les utilisateurs devaient indiquer, pour chacune des 7 journées constituant la croisière, leur intention de participer ou non à une certaine activité. Autrement dit, ils avaient à organiser leurs activités sous la forme de planning journaliers. Nous avons ainsi collecté 23 contributions. Les caractéristiques du jeu de données sont décrites dans le Tab. 4.1.

4.2. *Evaluation*

Nous avons divisé les itinéraires créés par des utilisateurs (sondés) en deux parties : la première partie constitue les données historiques utilisées pour l'extraction des motifs de comportement d'utilisateurs et l'estimation des scores d'intérêts, tandis que la deuxième partie constitue le jeu de test pour évaluer notre approche. Des partitions de taille différente en terme du nombre de jours (de 1 à 6) ont été considérées. L'évaluation a été réalisée en deux étapes.

Dans un premier temps, nous avons évalué la précision de l'estimation des scores d'activités selon les métriques usuelles suivantes (Shani et Gunawardana, 2011) : Moyenne des erreurs moyennes (MAE), Racine carrée de la moyenne des erreurs des moindres carrés (RMSE), Précision au rang k et AUC (l'aire sous la courbe ROC). Nous avons défini k pour un utilisateur, comme étant égal au nombre moyen d'activités effectuées par l'utilisateur dans le passé. La motivation derrière cet ajustement se base sur le fait que la densité des activités effectuées par utilisateur varie d'un utilisateur à un autre. Nous avons considéré dix cas selon les méthodes de calcul des scores personnalisés, c'est-à-dire pour chacune des deux stratégies : (1) le score catégoriel, (2) le score textuel (basé sur le contenu), (3) le score temporel, (4) le score hybride, (5) le score basé sur le régression logistique (voir la Section 3.1). Les données binaires de participation des utilisateurs aux activités constituent notre vérité terrain.

Dans un deuxième temps, nous avons évalué la construction d'itinéraires en comparant ILS_TP et ILS aux séquences d'activités annotées par des utilisateurs (notées *"vérité terrain"*). Nous avons considéré la métrique que nous nommons "Similarité" qui représente le ratio entre les activités recommandées faisant partie de la séquence "vérité terrain" et la longueur de la séquence " vérité terrain".

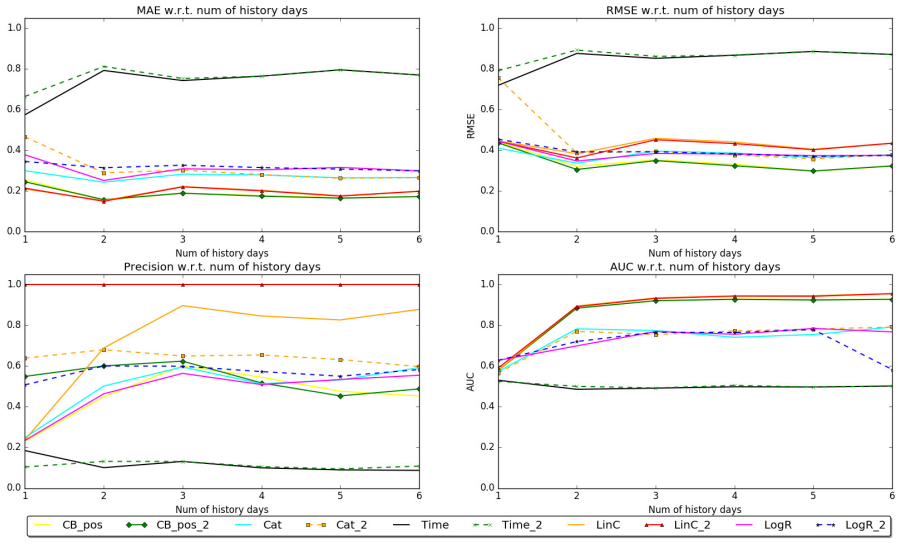

Figure 1. *Résultats d'évaluation de l'estimation des scores personnalisés d'activités.*

5. Résultats et discussion

Nous avons implémenté l'estimation des scores d'activités et de la probabilité de transition en Python 3.5.2, et l'algorithme de construction d'itinéraires en GNU Octave. Tout d'abord, nous avons évalué la précision des scores obtenus par notre méthode (voir Fig. 1). Nous constatons que l'utilisation de plusieurs facteurs améliore la puissance prédictive de notre modèle. Ainsi, les scores hybrides (noté *Hybrid*) obtiennent les meilleurs résultats en terme de précision et AUC, en cédant la première place aux scores basés sur le contenu (notés *CB_pos* et *CB_pos_2*) en terme de MAE. Nous observons que la stratégie 2 (lignes pointillées sur la figure) est la plus performante. Ces résultats étaient attendus vu le concept d'enrichissement des données historiques mis en place dans la Stratégie 2. Enfin, nous constatons que la variation du nombre de jours historiques pris en compte n'influence pas la performance de l'approche au dela de 2 jours. Les scores proposés, en particulier le score hybride intégré à la stratégie 2, sont de bons estimateurs de l'intérêt des utilisateurs pour une activité.

Pour la suite de l'évaluation, nous avons considéré les scores hybrides calculés dans le cadre des stratégies 1 et 2 comme des données d'entrée pour les algorithmes ILS et ILS_TP. Nous avons fait varier le nombre de jours d'historique de 1 à 6. Les résultats montrent que l'utilisation de la probabilité de transition pour la construction de l'itinéraire améliore la performance. Le taux moyen d'amélioration d'ILS_TP par rapport ILS est de 7.3% en utilisant le score hybride de la stratégie 1 et de 14.1% en utilisant celui de la stratégie 2 (voir Tab. 6). Nous constatons une différence des

Tableau 6. *Amélioration d'ILS_TP par rapport ILS en terme de similarité (précision) par rapport à la vérité terrain, %*

Jours d'historique	1	2	3	4	5	6	Average
Stratégie 1	6.1	3.4	6.5	6.2	9.9	11.3	7.4
Stratégie 2	25.0	10.3	4.4	13.5	14.3	11.3	13.5

résultats en fonction du nombre de jours historiques, qui s'explique par l'utilisation des données historiques de l'utilisateur afin de calculer la probabilité de transition.

6. Conclusion

Dans cet article, nous avons formalisé et étudié le problème de recommandation de séquences d'activités lors d'évènements distribués. Ce problème représente un vrai défi en combinant à la fois le problème de recommendation des activités uniques et le problème de création d'itinéraire personnalisé. Nous avons proposé une approche intégrée de recommandation de séquence d'activités qui explore trois critères pour estimer des scores de similarité entre le profil de l'utilisateur et une activité : 1-les catégories des activités, 2-leurs descriptifs et 3-les préférences temporelles de l'utilisateur. Notre approche prend en compte l'aspect séquentiel du comportement de l'utilisateur afin de lui recommander un meilleur planning d'activités. Nous avons suggéré deux stratégies pour estimer des scores d'intérêt. Nous avons évalué notre approche sur un jeu de données que nous avons créé. Les expérimentations ont démontré que notre solution permet d'obtenir de meilleurs résultats que l'algorithme de l'état-de-l'art, ILS.

Notre approche permet de réduire l'impact des incertitudes provenant des préférences incertaines et des relations imprécises entre les caractéristiques des activités et les intérêts des utilisateurs. Notons que le fait qu'un utilisateur soit intéressé par une activité n'aboutit pas forcément à sa participation à cette activité et inversement. Ainsi, selon notre étude utilisateur, 15.73% d'activités auxquelles les utilisateurs ont choisi de participer n'avaient pas été marquées comme intéressantes, tandis que 58.12% des activités marquées comme intéressantes n'ont pas été effectuées par l'utilisateur.

Dans les travaux futurs, nous envisageons d'explorer différentes directions de recherche. Premièrement, d'autres types de critères peuvent être pris en compte pour l'estimation des scores personnalisés des activités, comme par exemple son age, son genre, ou le groupe avec lequel l'utilisateur participe à une activité. Deuxièmement, nous souhaitons développer la phase de construction d'itinéraire en prenant en compte de multiples fenêtres temporelles et de multiples localisations des activités. Finalement, nous souhaitons utiliser le crowdsourcing pour l'évaluation, en poursuivant trois objectifs : 1. Annoter l'ensemble des activités selon les intérêts de l'utilisateur ; 2. Créer des séquences d'activités ; 3. Évaluer les résultats de recommandation.

Remerciements

Diana Nurbakova est soutenue par la Région Auvergne Rhône Alpes.

7. Bibliographie

Fonteles A. S., Bouveret S., Gensel J., *Web and Wireless Geographical Information Systems : 14th International Symposium, W2GIS 2015, Proceedings*, chapter Opportunistic Trajectory Recommendation for Task Accomplishment in Crowdsourcing Systems, p. 178-190, 2015.

Gavalas D., Konstantopoulos C., Mastakas K., Pantziou G., « Mobile recommender systems in tourism », *J. Netw. Comput. Appl.*, vol. 39, p. 319 - 333, 2014.

Li X., Cong G., Li X.-L., Pham T.-A. N., Krishnaswamy S., « Rank-GeoFM : A Ranking Based Geographical Factorization Method for Point of Interest Recommendation », *Proceedings of the 38th ACM SIGIR*, p. 433-442, 2015.

Macedo A., Marinho L., Santos R., « Context-Aware Event Recommendation in Event-based Social Networks », *Proceedings of the 9th ACM RecSys*, p. 123-130, 2015.

Nurbakova D., Laporte L., Calabretto S., Gensel J., « Recommendation of Short-Term Activity Sequences During Distributed Events », *Procedia Computer Science*, vol. 108, p. 2069 - 2078, 2017. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

Sang J., Mei T., Xu C., « Activity Sensor : Check-In Usage Mining for Local Recommendation », *ACM Trans. Intell. Syst. Technol.*, vol. 6, n° 3, p. 41 :1-41 :24, April, 2015.

Schaller R., Harvey M., Elsweiler D., « RecSys for Distributed Events : Investigating the Influence of Recommendations on Visitor Plans », *Proceedings of the 36th ACM SIGIR*, p. 953-956, 2013.

Shani G., Gunawardana A., *Recommender Systems Handbook*, chapter Evaluating Recommendation Systems, p. 257-297, 2011.

Tchernykh A., Schwiegelsohn U., Alexandrov V., ghazali Talbi E., « Towards Understanding Uncertainty in Cloud Computing Resource Provisioning », *Procedia Computer Science*, vol. 51, p. 1772 - 1781, 2015. International Conference On Computational Science, {ICCS} 2015 Computational Science at the Gates of Nature.

Vansteenwegen P., Souffriau W., Oudheusden D. V., « The orienteering problem : A survey », *Eur. J. Oper. Res.*, vol. 209, n° 1, p. 1 - 10, 2011.

Vansteenwegen P., Souffriau W., Vanden B. G., Van Oudheusden D., « Iterated Local Search for the Team Orienteering Problem with Time Windows », *Comput. Oper. Res.*, vol. 36, n° 12, p. 3281-3290, December, 2009.

Zhang J.-D., Chow C.-Y., « GeoSoCa : Exploiting Geographical, Social and Categorical Correlations for Point-of-Interest Recommendations », *Proceedings of the 38th ACM SIGIR*, p. 443-452, 2015a.

Zhang J.-D., Chow C.-Y., « Spatiotemporal Sequential Influence Modeling for Location Recommendations : A Gravity-based Approach », *ACM Trans. Intell. Syst. Technol.*, vol. 7, n° 1, p. 11 :1-11 :25, 2015b.

Zhang J.-D., Chow C.-Y., Li Y., « LORE : Exploiting Sequential Influence for Location Recommendations », *Proceedings of the 22Nd ACM SIGSPATIAL*, p. 103-112, 2014.