ALGeoSPF: Un modèle de factorisation basé sur du clustering géographique pour la recommandation de POI

Jean-Benoît Griesner*, Talel Abdessalem*,**
Hubert Naacke*** Pierre Dosne*

*LTCI, Télécom ParisTech
Paris, France
griesner@telecom-paristech.fr,

**UMI CNRS IPAL, National University of Singapore
talel.abdessalem@telecom-paristech.fr

***Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606
Hubert.Naacke@lip6.fr

Résumé. La recommandation de points d'intérêts est devenue une caractéristique essentielle des réseaux sociaux géo-localisés qui a accompagné l'émergence des échanges massifs de données digitales. Cependant les faibles densités de points d'intérêts visités par les utilisateurs rendent le problème difficile à traiter, d'autant plus que les espaces de mobilité des utilisateurs sont très hétérogènes, allant de la ville au monde entier. Dans ce papier nous explorons l'impact d'une approche de clustering spatial sur la qualité de la recommandation. Notre approche est basée sur un modèle de factorisation de matrices de Poisson et un réseau social inféré des différents comportements de mobilité. Nous avons conduit une évaluation comparative des performances de notre approche sur un jeu de données réaliste. Les résultats expérimentaux montrent que notre approche permet une précision supérieure aux techniques de recommandation alternatives.

1 Introduction

La recommandation de *points d'intérêts* (ou POI pour *Point Of Interest*) consiste à proposer à un utilisateur une liste de POI qu'il pourrait être intéressé de visiter. Ce problème est devenu une composante majeure des réseaux sociaux géo-localisés (ou LBSN pour *Location-Based Social Networks*) permettant aux utilisateurs de découvrir de nouveaux POI mais également aux POI d'augmenter leur attractivité. Il faut toutefois tenir compte de différentes caractéristiques des LBSN, à savoir : (i) **sparsité :** la densité de la matrice utilisateur-POI est très faible, (ii) **fréquences de visite** : nous savons seulement combien de fois un utilisateur a été localisé dans un lieu, et (iii) **feedback implicite** : nous ne disposons pas d'une évaluation explicite du POI par l'utilisateur.

Nous considérons ici le cas où les visites des utilisateurs couvrent une large étendue géographique pouvant aller jusqu'au monde entier. Quand la surface couverte par les données augmente les données deviennent de plus en plus creuses car le nombre de POI augmente tandis que le nombre moyen de POI visités par utilisateur reste constant. Il résulte que la densité (*i.e.* , la proportion de POI visités par utilisateur) diminue. Pour résoudre ce problème de **faible densité** nous avons étudié des solutions pour augmenter la densité des données sans réduire la zone géographique couverte. Notre idée est de définir un ensemble de **superPOI** chacun représentant un groupe de POI. L'ensemble des superPOI constitue une structure hiérarchique incluant d'autres superPOI. Le second problème à résoudre est de tenir compte des fluctuations des **comportements de mobilité** des utilisateurs, pouvant aller d'une échelle géographique locale (une ville) à globale (le monde). Notre principale motivation est de nous assurer que notre modèle de recommandation disposera de suffisamment d'informations pour chaque utilisateur. Notre approche ne se limite pas à 2 classes et peut s'adapter à un nombre de classes quelconque en fonction des données.

Ainsi les contraintes de densité voudraient que nous agrégions les POI dans des superPOI, et en même temps les comportements de mobilité exigent de conserver suffisamment de POI locaux. Pour trouver le bon **compromis** nous proposons une solution unifiée qui s'appuie sur notre définition de superPOI de façon à agréger les POI de manière flexible (*c.f.* section 5) et également qui classifie les utilisateurs de façon à trouver le meilleur niveau d'agrégation (*c.f.* section 5). Nous pouvons résumer les contributions auxquelles nous avons abouti ainsi :

— Une approche probabiliste de factorisation qui **passe à l'échelle**.
— Une **structure hiérarchique** pour définir plusieurs niveaux de **superPOI**.
— La définition de différents **comportements de mobilité** des utilisateurs.
— Des résultats expérimentaux qui confirment **l'efficacité** de notre approche sur un jeu de données de grandes dimensions.

2 État de l'art

Ye et al. (2011) ont proposé une méthode de filtrage collaboratif (CF) *memory-based* qui intègre les influences sociale et géographique. En particulier ils ont mis en évidence que le facteur géographique jouait un rôle essentiel dans la qualité du modèle. Malheureusement cette approche n'est pas capable de traiter des jeux de données à grande échelle. De nombreuses méthodes CF *model-based* ont aussi été proposées notamment par Zhang et Wang (2015); Hu et al. (2008). Parmi ceux-ci la classe de méthodes la plus exploitée est la classe des modèles de factorisation de matrice tels que présentés par Zhang et Wang (2015); Griesner et al. (2015). La factorisation de matrice vise à exprimer l'interaction utilisateur-objet en grâce à des vecteurs d'attributs latents. La factorisation de matrice probabiliste de Salakhutdinov et Mnih (2007) (PMF) est une approche qui vise à minimiser une fonction objectif des moindres carré avec des termes de régularisation quadratiques. Plus récemment Griesner et al. (2015) ont proposé un modèle pour traiter le problème de l'évaluation implicite. Leurs résultats étaient prometteurs mais la complexité de leur approche empêchait cette méthode de passer à l'échelle.

L'un des principaux problèmes est l'absence d'évaluation explicite de l'expérience utilisateur. Pour traiter les jeux de données implicites un modèle a été proposé il y a quelques années par Hu et al. (2008). Ce modèle distingue les préférences de l'utilisateur et la confiance que nous pouvons avoir dans cette préférence. Les auteurs ont démontré que leur approche était plus efficace que d'autres méthodes existantes sur de petits jeux de données. Cette méthode a été exploitée et enrichie par Lian et al. (2014) pour intégrer l'influence géographique des POI en modélisant le phénomène de clustering spatial directement dans le processus de factorisation. Cependant la complexité est trop élevée pour l'appliquer sur des jeux de données réels. La factorisation de Poisson a été proposée récemment par Gopalan et al. (2013) comme une solution alternative efficace. Il s'agit d'un modèle de factorisation probabiliste qui surpasse les modèles de factorisation alternatifs soumis aux mêmes contraintes de densité. De nombreux travaux récents ont proposé d'intégrer l'influence sociale tel que Zhang et Wang (2015). L'idée est d'exploiter les opinions des voisins d'un utilisateur sur des POI non visités par celui-ci. En particulier Zhang et Wang (2015) ont proposé un modèle appelé LTSCR qui utilise les similarités sociales des utilisateurs et les intègre dans un modèle de factorisation. Cependant les informations sociales ne sont généralement pas présentes dans les jeux issus des LBSN par souci de confidentialité.

3 Modèle des données

La plupart des LBSN disposent d'un ensemble d'utilisateurs $\mathcal{U}$, d'un ensemble de POI $\mathcal{P}$, d'informations temporelles $\mathcal{T}$ et sociales $\mathcal{S}$. Dans ce contexte un utilisateur u peut faire des *check-ins* (*i.e.* des visites) à un POI donné p à l'instant t.

Definition 1 *(**POI**) Un point d'intérêt correspond à un site unique associé à une activité spécifique (e.g. musée, restaurant, université...). Nous notons $\mathcal{P}$ l'ensemble des POI. Ici nous supposons que nous n'avons accès qu'aux localisations, i.e. , aux paires $(latitude,longitude)$ pour chaque POI.*

Definition 2 *(**Check-in**) Le check-in de l'utilisateur u visitant un POI p à l'instant t est associé au tuple $<u,p,t>$. Ainsi pour calculer la fréquence de visite de u à p il suffit d'agréger le nombre de check-ins correspondant. Etant donné que chaque POI est associé au moins à un superPOI, chaque check-in incrémente également la fréquence de visite de ce superPOI.*

Definition 3 *(**SuperPOI**) L'agrégation de plusieurs POI ou superPOI sur une zone géographique unique constitue un superPOI. L'ensemble des superPOI est un ensemble de POI ou superPOI. Les zones de chaque superPOI sont disjointes deux à deux (c.f. section 5).*

Definition 4 *(**Profil utilisateur**) Chaque utilisateur dispose d'un profil qui correspond à l'ensemble de tous ses check-ins : $\mathcal{P}^u = \{<u,p_i,t_j> \in \mathcal{D}\}$. L'agrégation de tous les profils utilisateur constitue le jeu de donnée complet $\mathcal{D} = \{\mathcal{P}^u / u \in \mathcal{U}\}$.*

Notre objectif est de recommander une liste de POI à un utilisateur donné en se basant sur ses check-ins passés et d'autres informations annexes. Ces informations annexes correspondent pour nous aux localisations. Nous utilisons ces localisations notamment pour construire un graphe d'accessibilité géographique implicite (AGRA).

Definition 5 *(**AGRA**) Notre graphe d'accessibilité AGRA, noté* $\mathbf{G}=(V,E,\rho)$ *est un graphe orienté où chaque noeud* $v \in V$ *représente un POI associé à ses coordonnées géographiques, chaque arête* $e=(p_i,p_j) \in E$ *n'existe que si la transition* $p_i \rightarrow p_j$ *est observée dans au moins un itinéraire utilisateur et* ρ *est une fonction qui associe à chaque arête* $e=(p_i,p_j)$ *son accessibilité correspondante* $\mathcal{A}_{i,j}$ *(définie à l'équation 3).*

Problem 1 ***Recommandation de POI à large échelle :*** *Étant donnée une collection de check-ins* $\mathcal{D}$ *de faible densité distribuée sur le monde entier, l'objectif est de fournir à un utilisateur* u *une top-k liste de POI qu'il pourrait avoir envie de visiter.*

4 Modèle des influences géographique et sociale

Cette section présente GeoSPF (*i.e.* , *Geographical Social Poisson Factorization*), notre méthode pour extraire les influences sociales implicites à partir des comportements de mobilité géographique. À cet effet nous introduisons notre graphe d'accessibilité (AGRA) que nous utilisons pour notre modèle géographique.

4.1 Idée générale

GeoSPF est basé sur l'hypothèse que le choix de l'utilisateur dépend d'une combinaison de préférences géographique, sociale et personnelle. Si nous notons $\alpha(u,p)$ le degré d'intérêt qu'un utilisateur u a pour un POI p, $\mathcal{S}(u,p)$ l'influence sociale que u a obtenu pour p, et $\mathcal{G}(u,p)$ la préférence géographique de l'utilisateur u concernant p, la probabilité d'observer la paire (u,p) dans les données devrait être directement proportionnelle à l'intérêt de u pour p, et diminuer de façon monotone, tel que :

$$\mathrm{P}(u,p) \propto \mathbb{F}[\alpha(u,p),\mathcal{G}(u,p),\mathcal{S}(u,p)] \tag{1}$$

où $\mathbb{F}[\cdot]$ est une fonction qui combine les intérêts personnels, l'influence sociale et l'influence géographique. Les approches existantes proposées par Lian et al. (2014); Griesner et al. (2015) ont vérifié que l'influence géographique a un impact significatif sur la qualité de la recommandation. Cependant ils utilisent en général un espace isotrope uniforme et exploitent seulement les distances entre les check-ins. Ainsi de telles approches ne tiennent pas compte des contraintes naturelles qui pourraient rendre la mobilité entre deux POI difficile même s'ils sont proches les uns des autres. Au-delà des distances, nous introduisons ici le concept d'accessibilité, comme nous le verrons dans la section 4.2, pour mieux intégrer l'influence géographique dans les choix des utilisateurs. Les **étapes principales** de GeoSPF sont les suivantes : (**i**) Nous construisons un graphe d'accessibilité AGRA basé sur les transitions observées (d'un POI à l'autre) et leurs probabilités, puis (**ii**) nous construisons un réseau social implicite ISN à partir de AGRA et des similitudes entre les historiques de transitions des utilisateurs, et enfin (**iii**) nous intégrons l'ISN dans un modèle de recommandation de factorisation sociale de Poisson pour obtenir notre modèle GeoSPF.

4.2 Accessibilité géographique

L'idée de l'accessibilité consiste à modéliser la probabilité qu'un utilisateur se déplace vers un POI p_{j+1} après avoir visité le POI p_j. Pour ce faire nous appliquons un modèle de Markov de premier ordre. Une transition est observée dans l'itinéraire d'un utilisateur u s'il existe dans le jeu de données deux check-ins consécutifs $<u,p_i,t_1>$ et $<u,p_j,t_2>$ effectués dans deux POI différents p_i et p_j à deux timestamps t_1 et t_2, tels que $t_1<t_2$ et si aucun autre enregistrement intermédiaire $<u,p_k,t'>$ ($t_1<t'<t_2$) ne se trouve dans le jeu de données. Nous notons cette transition comme suit : $p_i \rightarrow p_j$ dans le reste de l'article. Ainsi pour un utilisateur donné la probabilité de visiter p_{j+1} sera déduite de sa dernière visite. Plus formellement nous avons $P(p_{j+1}|p_j,p_{j-1},...,p_1)=P(p_{j+1}|p_j)$ où nous définissons $P(p_{j+1}|p_j)$ comme probabilité de transition $\mathcal{T}_{j,j+1}$ de p_j à p_{j+1}. Nous pouvons calculer cette probabilité en utilisant l'estimation empirique du maximum de vraisemblance comme suit :

$$\mathcal{T}_{j,j+1}=P(p_{j+1}|p_j)=\frac{N(p_j,p_{j+1})}{N(p_j)} \tag{2}$$

où $N(p_j,p_{j+1})$ est le nombre d'utilisateurs ayant la séquence $p_j \rightarrow p_{j+1}$ dans leur profil, et $N(p_j)$ est le nombre d'utilisateurs ayant visité p_j. Nous avons $N(p_j,p_{j+1}) \leqslant N(p_j)$, donc nous savons que $\mathcal{T}_{j,j+1}$ est borné : $\mathcal{T}_{j,j+1} \in [0,1]$. Observons que pour calculer cette probabilité les check-ins doivent suivre l'ordre chronologique. Ensuite nous pouvons combiner cette probabilité avec l'information géographique afin d'estimer l'accessibilité $\mathcal{A}_{j,j+1}$ entre les POI p_j et p_{j+1}. Nous définissons cette accessibilité comme suit :

$$\mathcal{A}_{j,j+1}=\frac{1}{0.5+d(p_j,p_{j+1})} \cdot \mathcal{T}_{j,j+1} \tag{3}$$

où $\mathcal{T}_{j,j+1}$ fait référence à l'équation 2 et $d(p_j,p_{j+1})$ est la distance euclidienne entre les POI p_j et p_{j+1}. Si p_{j+1} est loin de p_j l'accessibilité sera faible. Mais lorsque beaucoup de transitions sont observées de p_j à p_{j+1} l'accessibilité augmente. L'équation 3 est inspirée des poids géographiques utilisés par Liu et Xiong (2013). La valeur de 0,5 au dénominateur est déterminée empiriquement en fonction des spécificités des données. Cela signifie que nous accordons plus d'importance aux distances inférieures à 500 mètres. Nous utiliserons ensuite cette accessibilité pour définir notre *Jaccard symétrique pondéré d'accessibilité* dans la sous-section 4.3.

4.3 AGRA : Graphe d'accessibilité

Des travaux précédents tels que Ma et al. (2011); A. et al. (2014) ont montré que les influences sociales jouaient un rôle important dans la qualité finale de la recommandation de POI. Cependant en général dans les LBSN nous n'avons pas accès à un réseau social explicite : nous avons seulement accès à l'historique des check-ins. Ainsi notre approche vise à construire un réseau social implicite (*implicit social network* ou ISN) basé sur la similarité entre les profils des utilisateurs et leurs transitions dans le graphe AGRA. Nous définissons ci-dessous quatre mesures de similarité possibles choisies pour leur flexibilité et la qualité de leurs résultats.

Adamic/Adar : Cette mesure accorde une grande importance aux transitions rares (c'est-à-dire avec une faible accessibilité). Intuitivement plus les deux utilisateurs partagent des POI impliqués dans des transitions rares, plus on s'attend à ce qu'ils soient proches. Ainsi avec $D(\cdot)$ la fonction donnant le degré d'un noeud, nous définissons :

$$S_{\mathrm{AA}}(u_1,u_2)=\sum_{v\in\mathcal{P}^{u_1}\cap\mathcal{P}^{u_2}}\frac{1}{\log(D(v))} \tag{4}$$

Jaccard standard : Il s'agit de la mesure de Jaccard standard. Ainsi nous définissons la similarité de Jaccard $S_J(u_1,u_2)$ de deux utilisateurs u_1 et u_2 comme suit :

$$S_{\mathrm{J}}(u_1,u_2)=|\mathcal{P}^{u_1}\cap\mathcal{P}^{u_2}|/|\mathcal{P}^{u_1}\cup\mathcal{P}^{u_2}| \tag{5}$$

Jaccard pondéré symétrique : Avec cette mesure nous étendons la mesure Jaccard standard en considérant l'accessibilité entre les points d'intérêt visités. Pour ce faire nous ajoutons à l'ensemble des POI visités ceux ($\Gamma(\mathcal{P}^u)$) qui sont accessibles en un seul bond en parcourant AGRA. Soit $G=\Gamma(\mathcal{P}^{u_1})\cup\Gamma(\mathcal{P}^{u_2})$ l'ensemble des POI visités par u_1 ou u_2. Soit $N=|G|$. Soient $\rho^{u_1}\in\mathbb{R}_+^N$ et $\rho^{u_2}\in\mathbb{R}_+^N$ deux vecteurs de poids d'accessibilité. Le vecteur ρ^{u_1} est construit ainsi : $\forall i\in[0,N]$ **Si** $p_i\in\mathcal{P}^{u_1}$ **Alors** $\rho_i^{u_1}=1$ **Sinon Si** $p_i\in\Gamma(\mathcal{P}^{u_1})$ **Alors** $\rho_i^{u_1}=\sum_{v\in\mathcal{P}^{u_1}}\mathcal{A}_{v,p}$ **Sinon** $\rho_i^{u_1}=0$. De même nous construisons le vecteur ρ^{u_2}. Il s'agit d'une métrique symétrique. Ainsi nous définissons la similarité de Jaccard symétrique pondérée par l'accessibilité $S_{AWS}(u_1,u_2)$ comme suit :

$$S_{AWS}(u_1,u_2)=\frac{\sum_{i\in[0,N]}\min(\rho_i^{u_1},\rho_i^{u_2})}{\sum_{i\in[0,N]}\max(\rho_i^{u_1},\rho_i^{u_2})} \tag{6}$$

Jaccard pondéré antisymétrique : Dans cette métrique nous essayons de prendre en compte l'asymétrie d'influence qui pourrait exister entre deux utilisateurs. Pour ce faire nous changeons la définition de G comme suit : $G=\Gamma(\mathcal{P}^{u_1})\cup\mathcal{P}^{u_2}$. Au lieu d'étendre les deux ensembles $\mathcal{P}^{u_1}$ et $\mathcal{P}^{u_2}$ nous étendons seulement l'ensemble des POI visités par l'utilisateur u_1. Ensuite nous calculons le Jaccard pondéré antisymétrique $S_{AWA}(u_1,u_2)$ en utilisant l'équation 6. Notons que $S_{AWA}(u_1,u_2)\neq S_{AWA}(u_2,u_1)$.

4.4 GeoSPF : Modèle de factorisation

La factorisation de Poisson a été proposée par Gopalan et al. (2013); Chaney et al. (2015). Le but des approches de factorisation est d'approximer la matrice utilisateur-POI $\mathbf{X}$ représentant les fréquences de visite, par le produit scalaire de facteurs latents tels que : $\mathbf{X}\approx\mathbf{U}\mathbf{V}^T$, où $\mathbf{U}\in\mathbb{R}^{m\times k}$ et $\mathbf{V}\in\mathbb{R}^{n\times k}$ avec $k\ll\min(m,n)$. La factorisation de Poisson (PF) est une approche générative probabiliste basée sur une loi de Poisson pour modéliser les observations. Il s'agit d'un modèle rapide et adapté aux données creuses. Récemment Chaney et al. (2015) en ont proposé une extension appelé SPF (*Social Poisson Factorization*). L'extension SPF aboutit à de bons résultats pour la recommandation de POI car SPF possède des propriétés intéressantes qui correspondent à nos besoins en termes de **qualité** et de **passage à l'échelle**. De plus SPF permet d'intégrer **l'information sociale** qui est importante dans notre contexte. Enfin SPF sépare les

questions : *qui est membre du cercle* ? et *quelle influence ce membre transmet-il réellement* ? SPF suppose que l'appartenance au cercle est connue à l'avance alors que le degré d'influence est appris. Cette **séparation** est essentielle ici car le degré d'influence d'un utilisateur ne dépend pas des POI qu'il partage avec les autres utilisateurs mais plutôt des interactions cachées (non divulguées) que les utilisateurs peuvent avoir. GeoSPF s'appuie sur la distribution suivante :

$$y_{i,j} \sim \text{Poisson}\left[\mathbf{u}_i^T \cdot \mathbf{v}_j + \sum_{k \in V(i)} \mathbf{s}_{i,k} \cdot x_{k,j}\right] \tag{7}$$

où $V(i)$ fait référence à l'ensemble des voisins de l'utilisateur i dans l'ISN, et $\mathbf{s}_{i,k}$ fait référence au facteur latent d'influence sociale. Cette variable modélise l'influence que le voisin k a sur l'utilisateur i. Le choix du voisinage $V(i)$ est important car $V(i)$ contiendra tous les voisins les plus proches de l'utilisateur i. Ainsi contrairement à SPF nous n'utilisons aucun réseau social explicite. De plus nous pouvons ajuster la qualité de la recommandation en fonction des métriques de similarité. Enfin nous pouvons également appliquer des filtres de seuillage soit sur le graphe d'accessibilité, soit sur le graphe social implicite.

5 Modèle des influences locale et globale

Notre modèle *Augmented Local-Global GeoSPF* (noté ALGeoSPF) consiste à définir des strates locales et globales de superPOI afin d'augmenter la densité des données. Le principal avantage d'ALGeoSPF est sa capacité à détecter différents comportements de mobilité, des échelles locale jusqu'à mondiale. Un autre avantage est qu'elle permet d'isoler toutes les étapes du processus de recommandation (*i.e.* , inférence, apprentissage, prédiction du réseau social) au sein de chaque classe d'utilisateurs. On évite ainsi toute propagation de bruit à travers les classes.

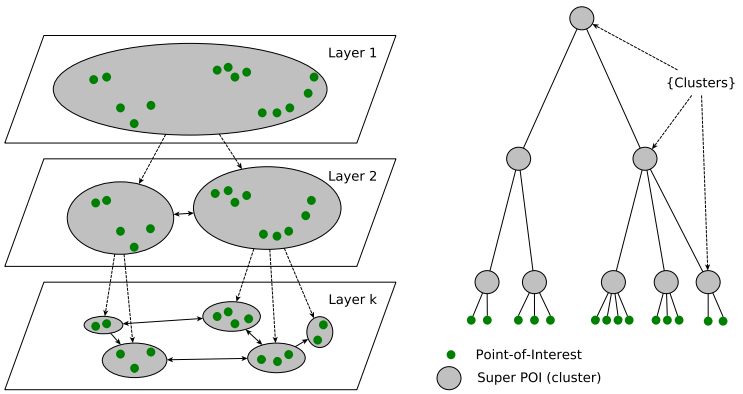


FIG. 1: Illustration de 3 strates hiérarchiques de check-ins et de superPOI associés.

5.1 Strates hiérarchiques de SuperPOI

Notre agrégation de SuperPOI permet de traiter des jeux de données de faible densité à grande échelle en agrégeant des portions des jeux de données d'origine. On note $\mathfrak{P}$ l'ensemble de toutes les strates de superPOI. Nous définissons une structure multi-strates pour agréger progressivement chaque POI en superPOI visité par un nombre croissant d'utilisateurs. La figure 1 représente une illustration de cette structure. Cela permet de répondre aux exigences de densité de notre problème pour différentes classes d'utilisateurs. Soit k le nombre de strates, $\mathfrak{P}_k$ l'ensemble des superPOI définis au niveau k et N^k_{max} le nombre maximum d'utilisateurs visitant un superPOI de $\mathfrak{P}_k$. La condition suivante définit la zone maximale qu'un superPOI représente, en notant p un superPOI : $\forall p \in \mathfrak{P}_k, N(p) < N^k_{max}$ avec $N(p)$ étant le nombre total d'utilisateurs distincts visitant le POI p. De plus chaque niveau vise à agréger les POI autant que possible pour s'assurer que chaque p n'est pas "trop petit", c'est-à-dire qu'il n'y a pas p' dans la strate supérieure $\mathfrak{P}_{k+1}$ tel que p' agrège p et p' satisfait $N(p') < N^k_{max}$.

5.2 Algorithme de clustering géographique

Nous adoptons une approche de clustering qui consiste à diviser l'espace géographique initial (par exemple le monde entier pour le jeu de données à grande échelle YFCC) en plusieurs cellules rectangulaires identiques. Les approches traditionnelles, comme proposé par Al-Ghossein et Abdessalem (2016), sont basées sur les *quad-tree* : il s'agit de construire un arbre où la racine est la carte du monde entier et chaque noeud est le quart de sa région parente. Nous divisons récursivement une cellule c jusqu'à ce qu'elle satisfasse la condition concernant le nombre d'utilisateurs différents qui ont fait des check-ins dans cette cellule : $N(c) < N_{max}$. Les cellules satisfaisant cette condition sont choisies pour être un superPOI. Le résultat de l'algorithme de clustering est un ensemble de *cellules superPOI* notées S. Le paramètre N_{max} permet de contrôler le niveau d'agrégation. Nous spécifions l'état d'une cellule superPOI basée sur $N(\cdot)$ au lieu du nombre de POI car ainsi on détecte mieux les comportements de mobilité des utilisateurs lorsque de nombreux POI populaires sont proches les uns des autres dans une zone comptant un petit nombre de POI, ce qui arrive relativement souvent. L'algorithme est présenté dans l'algorithme 1.

5.3 Sélection personnalisée de classe

Au moment où un utilisateur attend une recommandation, nous nous appuyons sur l'algorithme de clustering pour ajuster (c'est-à-dire pour optimiser) les deux paramètres qui affectent la densité du jeu de données : la cellule initiale et la taille du cluster. La **cellule initiale** est la cellule sur laquelle appliquer le clustering. La **taille du cluster** est définie par N_{max}. Lorsque N_{max} augmente les superPOI seront plus vastes mais il y en aura moins, ce qui augmentera la densité du jeu de données. Cependant N_{max} est borné : pour chaque utilisateur il existe un maximum N_{max} (noté N^{user}_{max}) au-delà duquel la recommandation n'est plus possible car l'utilisateur n'aura pas visité suffisamment de superPOI distincts.

Algorithm 1 Méthode de clustering *Top-down* pour ALGeoSPF

Input :
— N_{max} : nombre maximum d'utilisateurs ayant visité une cellule.
Global Output :
— S : l'ensemble des superPOI.
Initialize : $S \leftarrow \emptyset$
function WORLDTOSUPERPOIS (C : une cellule)
    Split C en 4 cellules rectangulaires identiques $C_1, \cdots, C_4$
    for each C_i **do**
        if $N(C_i) > N_{max}$ et $\#POIs(C_i) \geqslant 2$ **then**
            worldToSuperPOIs (C_i)
        else Mettre C_i dans S

En réglant ces paramètres nous avons constaté que leur impact sur la densité varie beaucoup selon les utilisateurs, ce qui justifie notre approche personnalisée. Pour certains utilisateurs une cellule initiale plutôt petite (*e.g.* *Paris*) donne une densité élevée. Pour d'autres, bien que *Paris* soit la cellule initiale (car la cellule *Paris* contient tous les check-ins de l'utilisateur), la cellule initiale *France* permet d'avoir N_{max} et permet une densité plus élevée. Plus généralement nous avons observé qu'une telle méthode d'optimisation permet de détecter plusieurs classes d'utilisateurs partageant le même couple quasi-optimal de paramètres (cellule initiale, N_{max}). Nous observons que chaque utilisateur appartient à l'une ou l'autre de ces classes. Ainsi nous associons chaque utilisateur à son optimal personnalisé N_{max}^{user} qui est le N_{max} qui maximise la densité des check-ins distincts dans les clusters.

6 Évaluation expérimentale

Nous avons conduit des expériences sur trois jeux de données issus du monde réel contenant des check-ins provenant de LBSN largement utilisés, à savoir : `YFCC`, `Gowalla` et `Foursquare`. Afin d'évaluer la qualité à différentes échelles géographiques nous avons filtré les jeux de données de façon à ce qu'ils couvrent respectivement une petite, une moyenne et une grande superficie. Ainsi Gowalla@Paris couvre une ville, Foursquare couvre une région, Gowalla couvre un pays et YFCC couvre l'Europe. Le jeu de données YFCC a été récemment proposé par Thomee et al. (2016). La table 1 présente quelques statistiques sur les jeux de données utilisés. Dans notre protocole d'évaluation 20% des données sont sélectionnées au hasard pour les tests et le reste est utilisé pour l'apprentissage du modèle. Ainsi nous comparons [4] GeoSPF et ALGeoSPF avec les modèles de recommandation suivants : (i) **NMF** *Non Negative Matrix Factorization* de Lee et Seung (2000), (ii) **PMF** *Probabilistic Matrix Factorization*

1. données disponibles ici : https ://sites.google.com/site/yangdingqi/home/foursquare-dataset
2. données disponibles ici : http ://www.yongliu.org/datasets
3. données disponibles ici : https ://webscope.sandbox.yahoo.com/catalog.php ?datatype=i&did=67
4. Le code pour reproduire les expériences est disponible ici : https ://gitlab.telecom-paristech.fr/griesner/geopfModeles

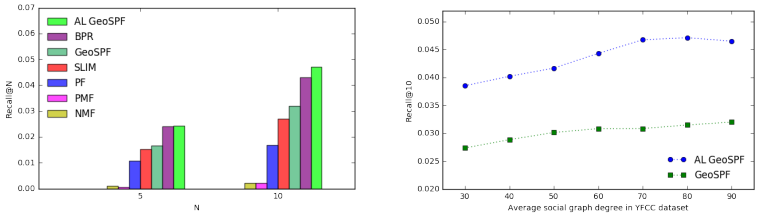
TAB. 1: Statistiques sur les jeux de données

Dataset	#Check-ins	#Users	#POIs	avg#POIs	Densité
Gowalla@Paris	42323	2384	4895	5.6	0.362 %
Foursquare [1]	109077	4825	19645	3.1	0.115 %
Gowalla [2]	191365	6749	24353	4.1	0.116 %
YFCC [3]	48453357	214328	12758657	61.2	0.0017 %

de Salakhutdinov et Mnih (2007), (iii) **SLIM** *Sparse Linear Methods* de Ning et Karypis (2012), (iv) **BPR** *Bayesian Personalized Ranking* de Rendle et al. (2009), (v) **WRMF** *Weighted Regularized Matrix Factorization* de Hu et al. (2008), qui donne de très bons résultats, (vi) **PoissonMF** de Gopalan et al. (2013) correspond au modèle probabiliste de Poisson qui sert de socle à notre approche, (vii) **GeoSPF** notre méthode, et enfin (iix) **ALGeoSPF** qui correspond à notre modèle final *Augmented Local-Global GeoSPF*.

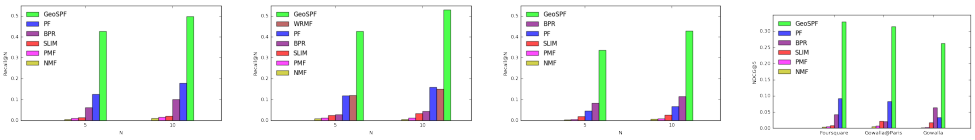
La figure 3 présente les performances globales obtenues par chaque méthode comparatives listées ci-dessus. Sur la figure 3a le Rappel@5 et le Rappel@10 sont présentés pour le jeu de données Foursquare. La figure 3b (resp. 3c) concerne Gowalla@Paris (resp. Gowalla). Enfin la figure 3d présente la métrique NDCG@5 sur les trois jeux de données. Conformément à nos prévisions, NMF et PMF donnent une qualité relativement mauvaise étant donné qu'ils ont été conçus pour opérer sur des jeux de données explicites. Cette observation est donc cohérente avec les résultats de Liu et Xiong (2013). De même SLIM ne parvient pas à fournir une qualité satisfaisante car il n'est opérant, lui aussi, que sur les jeux de données contenant une évaluation explicite de la part de l'utilisateur. Malheureusement la complexité de WRMF le rend pratiquement inutilisable sur de grands jeux de données : le temps de calcul de WRMF rend sont usage en situation réelle prohibitif. Comme résultat majeur de nos expériences nous observons que l'avantage relatif de GeoSPF sur tous les jeux de données est d'environ 200%. Ce gain notable rend GeoSPF approprié pour la recommandation de POI sur de vastes zones géographiques. Il confirme que l'exploitation d'informations contextuelles restreintes (uniquement le GPS et la date d'enregistrement) par le biais d'une solution géographique et sociale combinée donne une qualité finale élevée.

La figure 2 présente la qualité d'ALGeoSPF appliqué sur le jeu de données YFCC en considérant les utilisateurs urbains isolés des globetrotters. Plus précisément la figure 2a présente le rappel@10 de GeoSPF et d'ALGeoSPF pour différentes tailles moyennes du réseau social implicite. La figure 2b présente le rappel@5 et le rappel@10 des autres méthodes ainsi qu'ALGeoSPF sur le jeu de données YFCC (en utilisant une taille de réseau social fixe de 80). Nous observons qu'ALGeoSPF surpasse les autres méthodes bien que BPR soit très proche. Nous observons également que les mesures de rappel des modèles testés pour le jeu de données YFCC sont bien plus faibles que d'autres jeux de données, ce qui est dû à la faible densité des données, elle-même directement liée à l'étendue de la zone géographique couverte.



(a) Rappel@10 sur le YFCC (b) Rappel@5 et Rappel@10

FIG. 2: Rappel@10 sur la figure 2a de GeoSPF et d'ALGeoSPF en fonction du degré moyen du graphe social. Rappel@5 et Rappel@10 sur la figure 2b.



(a) Foursquare (b) Gowalla@Paris (c) Gowalla (d) Tous les datasets.

FIG. 3: Rappel@N sur les figures 3a, 3b et 3c. NDCG@5 sur la figure 3d.

7 Conclusions

Dans cet article nous avons proposé une nouvelle approche nommée ALGeoSPF qui passe à l'échelle pour la recommandation de POI dans les LBSN. L'objectif principal d'ALGeoSPF est de construire un modèle qui ne souffre pas de la faible densité des données des LBSN et qui prenne en compte les comportements de mobilité spécifiques des utilisateurs. Basés sur les concepts de *superPOI* et *d'accessibilité* que nous avons introduits, notre approche **(i)** construit efficacement un modèle de factorisation implicite large-échelle et **(ii)** intègre les préférences de mobilité de l'utilisateur dans une structure hiérarchique et enfin **(iii)** présente de meilleurs résultats que la plupart des approches existantes sur des jeux de données de grandes dimensions. Notamment nous avons démontré par des résultats expérimentaux qu'ALGeoSPF surpasse de façon significative toutes les approches alternatives testées en termes de *rappel* et de *NDCG*. Nous observons en particulier que nous sommes parmi les premiers à tester une approche de recommandation de POI sur le jeu de données YFCC.

Références

A., R., A. J., et C. J. Tauro (2014). A novel, generalized recommender system for social media using the collaborative-filtering technique. *SIGSOFT Softw. Eng. Notes 39*(3).

Al-Ghossein, M. et T. Abdessalem (2016). Somap : Dynamic clustering and ranking of geotagged posts. WWW '16 Companion, pp. 151–154.

Chaney, A. J., D. M. Blei, et T. Eliassi-Rad (2015). A probabilistic model for using social networks in personalized item recommendation. RecSys '15, pp. 43–50. ACM.

Gopalan, P., J. M. Hofman, et D. M. Blei (2013). Scalable recommendation with poisson factorization. *CoRR abs/1311.1704*.

Griesner, J., T. Abdessalem, et H. Naacke (2015). POI recommendation : Towards fused matrix factorization with geographical and temporal influences. pp. 301–304.

Hu, Y., Y. Koren, et C. Volinsky (2008). Collaborative filtering for implicit feedback datasets. ICDM '08.

Lee, D. D. et H. S. Seung (2000). Algorithms for non-negative matrix factorization. In *In NIPS*, pp. 556–562. MIT Press.

Lian, D., C. Zhao, X. Xie, G. Sun, E. Chen, et Y. Rui (2014). Geomf : Joint geographical modeling and matrix factorization for point-of-interest recommendation. KDD '14.

Liu, B. et H. Xiong (2013). Point-of-interest recommendation in location based social networks with topic and location awareness. ICDM'13.

Ma, H., T. C. Zhou, M. R. Lyu, et I. King (2011). Improving recommender systems by incorporating social contextual information. *ACM Trans. Inf. Syst. 29*(2), 9 :1–9 :23.

Ning, X. et G. Karypis (2012). Sparse linear methods with side information for top-n recommendations. RecSys '12, New York, NY, USA, pp. 155–162. ACM.

Rendle, S., C. Freudenthaler, Z. Gantner, et L. Schmidt-Thieme (2009). Bpr : Bayesian personalized ranking from implicit feedback. UAI '09, Arlington, Virginia, United States, pp. 452–461. AUAI Press.

Salakhutdinov, R. et A. Mnih (2007). Probabilistic matrix factorization. pp. 1257–1264. Curran Associates, Inc.

Thomee, B., D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, et L.-J. Li (2016). Yfcc100m : The new data in multimedia research. *Commun. ACM 59*(2).

Ye, M., P. Yin, W.-C. Lee, et D.-L. Lee (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. SIGIR '11.

Zhang, W. et J. Wang (2015). Location and time aware social collaborative retrieval for new successive point-of-interest recommendation. CIKM '15. ACM.

Summary

The task of points-of-interest recommendation has become an essential feature in social networks (LBSN) with the significant growth of shared data on LBSN. However it remains a challenging problem, because of the high level of sparsity of the data in LBSN. Moreover, in this context the mobility behavior of the users is very heterogeneous, ranging from urban to worldwide mobility. In this paper, we explore the impact of spatial clustering on the recommendation quality. The proposed approach combines spatial clustering with users' influences. It is based on a Poisson factorization model built on an implicit social network, inferred from the geographical mobility patterns. We conduct a comprehensive performance evaluation of our approach on the YFCC dataset (a very large-scale real-world dataset). The experiments show that our approach achieves a significantly superior quality compared to other existing recommendation techniques.