Un Modèle de Factorisation de Poisson pour la Recommandation de Points d'Intérêt

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

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

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

***UPMC Université Paris 06, LIP6, Paris, France
Hubert.Naacke@lip6.fr

Résumé. L'explosion des volumes de données circulant sur les réseaux sociaux géo-localisés (LBSN) rend possible l'extraction des préférences des utilisateurs. En particulier ces préférences peuvent être utilisées pour recommander à l'utilisateur des points d'intérêt en adéquation avec son profil. Aujourd'hui la recommandation de points d'intérêt est devenue une composante essentielle des LBSN. Malheureusement les méthodes de recommandation traditionnelles échouent à s'adapter aux contraintes propres aux LBSN, telles que la "sparsité" très élevée des données, ou prendre en compte l'influence géographique. Dans ce papier nous présentons un modèle de recommandation basée sur la factorisation de Poisson qui offre une solution efficace à ces contraintes. Nous avons testé notre modèle via des expérimentations sur un jeu de données réaliste issu du LBSN Foursquare. Ces expériences nous ont permis de démontrer une meilleure qualité de recommandation que 3 modèles de l'état-de-l'art.

1 Introduction

Les nombreux réseaux sociaux géolocalisés (ou "LBSNs" pour Location-Based Social Networks) tels que Foursquare, Flickr, Twitter etc. qui ont émergé ces dernières années permettent aux utilisateurs de partager leurs expériences concernant les Points d'Intérêt (ou POIs pour "Points Of Interest") qu'ils ont visités (i.e. les "checkins" de l'utilisateur). Le LBSN Flickr [1] par exemple compte plus de 110 millions d'utilisateurs

1. www.flickr.com

qui mettent en ligne un million d'images et de commentaires par jour. De tels volumes de données fournissent une information riche et précise, rendant possible de nouvelles formes de services en ligne, tels que la recommandation de POIs. La recommandation personnalisée de POIs est l'activité qui consiste à proposer à un utilisateur donné une liste de POIs qui soient susceptibles de l'intéresser. Aujourd'hui cette tâche est devenue une composante essentielle des LBSNs.

Malheureusement cette tâche reste un problème difficile. En effet la matrice initiale des checkins souffre d'une bien plus grande sparsité que les jeux de données traditionnels en recommandation. De surcroit la matrice ne contient que les fréquences de visite de chaque POI. Par conséquent on ne peut pas savoir si l'utilisateur apprécie ou non un POI. On parle dans ce cas de jeux de données avec "feedback" implicite Hu et al. (2008), Cheng et al.. Par ailleurs nous nous plaçons dans le contexte où seuls les checkins sont connus (i.e. <utilisateur, localisation, date>). La plupart des approches existantes négligent ces problèmes (sparsité, feedback implicite) en essayant uniquement d'adapter les modèles de recommandation traditionnels à cette problématique.

Dans ce papier nous proposons un modèle probabiliste de factorisation de Poisson pour la recommandation de POIs qui tienne compte de ces problèmes tout en passant à l'échelle. La suite de ce papier s'organise ainsi. La section 2 définit notre problème et décrit brièvement le modèle de factorisation de Poisson. La section 3 présente les deux approches de factorisation que nous proposons. Enfin nous présentons nos résultats expérimentaux dans la section 4, avant de conclure cet article dans la section 5.

2 Préliminaires

Définition du Problème Le but du modèle que nous proposons est de recommander une liste de POIs non visités à un utilisateur donné basé sur ses checkins passés. Ce problème nécessite deux choses : 1) modéliser les préférences utilisateurs de façon personnalisée ainsi que 2) la recommandation proprement dite à partir des données géographiques. Soit $U = \{u_1, u_2, ..., u_M\}$ un ensemble d'utilisateurs du LBSN, et soit $P = \{p_1, p_2, ..., p_N\}$ un ensemble de POIs, où chaque POI possède une localisation $l_j = (lon_j, lat_j)$, ainsi que des propriétés observables x_j (e.g. tags, titre, popularité...). Enfin soit c_{ij} le nombre de fois que l'utilisateur u_i a visité le POI p_j.

Factorisation de Poisson Notre modèle se base sur la factorisation de Poisson (PF). La PF est un modèle probabiliste de factorisation de matrices passant à l'échelle proposé récemment pour la recommandation par Gopalan et al. (2013). Dans ce modèle, si $c_{i,j}$ est le nombre de fois que l'utilisateur i a visité le POI j, la PF affirme que $c_{i,j}$ provient d'une distribution de Poisson, paramétrée par le produit scalaire des facteurs latents de l'utilisateur et du POI. La PF fournit ainsi l'estimation de $c_{i,j}$ selon la distribution suivante : $c_{i,j} \sim \text{Poisson}(\mathbf{u}_i^T \cdot \mathbf{v}_j)$ où $\mathbf{u}_i$ et $\mathbf{v}_j$ sont les vecteurs de facteurs latents de dimension K respectivement de l'utilisateur i et du POI j. A la différence de la PMF présentée plus haut, PF place des distributions a priori Gamma sur les facteurs latents de façon à s'adapter à la sparsité.

3 Recommandation Géographique de Poisson

Accessibilité Géographique : GeoPF La probabilité qu'un utilisateur se rende dans un POI donné ne dépend pas exclusivement de la distance qui l'en sépare. Ainsi pour quantifier cette probabilité, nous introduisons par la suite le concept d'accessibilité d'un POI à un autre. Pour ce faire nous utilisons un modèle de Markov d'ordre un. Dans ce modèle la probabilité de visiter le POI v_{j+1} sachant que l'on se trouve au POI v_j ne dépend que du POI v_j. Ainsi si nous définission $P(v_{j+1}|v_j)$ comme étant égale à la probabilité de transition de v_j à v_{j+1}, l'estimateur empirique de maximisation de vraisemblance donne pour $P(v_{j+1}|v_j)$ la valeur suivante : $P(v_{j+1}|v_j) = \frac{N(v_j,v_{j+1})}{N(v_j)}$ où $N(v_j, v_{j+1})$ correspond au nombre d'utilisateurs ayant fait la transition $v_j \rightarrow v_{j+1}$ dans le passé et où $N(v_j)$ est le nombre d'utilisateurs ayant visité v_j. Nous pouvons remarquer que nous utilisons ici l'information temporelle pour calculer la probabilité de transition. En effet ce calcul nécessite d'avoir ordonné auparavant les checkins par ordre chronologique A présent nous définissons l'accessibilité $\Phi(v_{j+1}, v_j)$ ainsi :

$$\Phi(v_{j+1}, v_j) = \frac{1}{0.5 + d(v_j, v_{j+1})} \cdot P(v_{j+1}|v_j) \tag{1}$$

où $P(v_{j+1}|v_j)$ est la probabilité de transition entre les POIs v_j et v_{j+1} calculée ci-dessus. $\Phi(v_{j+1}, v_j)$ utilise la distance euclidienne qui sépare les deux POIs et la probabilité de transitionner de l'un à l'autre. Ainsi plus $\Phi(v_{j+1}, v_j)$ sera élevée, plus la probabilité de visiter v_{j+1} sera importante. Nous utilisons l'accessibilité définie ci-dessus pour définir l'accessibilité moyenne d'un POI j pour un utilisateur i étant donné son itinéraire passé. Nous définissons finalement l'accessibilité moyenne du POI j pour un utilisateur i ainsi :

$$\widehat{\Phi}(i, j) = \frac{1}{N_i} \cdot \sum_{v_k \in L_i} \Phi(v_k, j) \tag{2}$$

où L_i est l'ensemble des POI déjà visités par l'utilisateur. Ainsi plus un POI sera accessible à partir des POIs précédents que l'utilisateur a l'habitude de visiter, plus $\widehat{\Phi}(i, j)$ sera élevée. Nous injectons ensuite cette accessibilité moyenne dans le modèle de recommandation de Poisson standard ainsi : $c_{i,j} \sim \text{Poisson}(\widehat{\Phi}(i, j) \cdot \mathbf{u}_i^T \cdot \mathbf{v}_j)$

Modèle Social : GeoSPF Notre second modèle repose sur SPF (pour "Social Poisson Factorisation") qui est un modèle de factorisation sociale de Poisson proposé récemment par Chaney et al. (2015). Ce modèle repose sur l'hypothèse que la matrice des checkins suit une loi de Poisson telle que : $c_{i,j} \sim \text{Poisson}(\mathbf{u}_i^T \cdot \mathbf{v}_j + \sum_{v \in N(i)} \mathbf{t}_{i,v} \cdot \mathbb{C}_{ij})$ où $N(i)$ est l'ensemble des amis de l'utilisateur i, et $\mathbf{t}_{i,v}$ est la variable aléatoire définissant l'influence sociale de l'ami v sur l'utilisateur i. Malheureusement la plupart des jeux de données ne contiennent pas de réseau social, rendant inutilisable ce modèle. C'est pourquoi nous proposons d'utiliser l'information géographique disponible pour construire un graphe social. Notre idée est de définir des distances entre utilisateurs basées sur la distance géographique des itinéraires de deux utilisateurs u_1, u_2. Nous

définissons cette distance géographique ainsi :

$$\text{Distance}(u_1, u_2) = \left[\frac{\sum_{i \in A} \text{distMin}(i, B)}{||A||} + \frac{\sum_{u \in B} \text{distMin}(i, A)}{||B||} \right] \cdot \frac{1}{2} \quad (3)$$

où A et B correspondent respectivement aux ensembles de POIs visités par u_1 et u_2, et où $\text{distMin}(i, X)$ est la distance minimale entre le POI i et n'importe quel POI appartenant à X.

4 Evaluation Expérimentale

Jeu de Données Pour tester la qualité de notre approche, nous avons utilisé un jeu de données issu de Foursquare [2]. Il contient 342850 checkins. Après avoir conservé uniquement les utilisateurs et les POIs qui avaient au minimum 5 checkins, il restait 194108 checkins faits par 2321 utilisateurs entre 5596 POIs. Nous avons vérifié que l'accessibilité (définie plus haut) ne dépendait pas exclusivement de la distance. Nous avons représenté sur la figure 1 la probabilité de transition en fonction de la distance. Nous observons que le jeu de données contient des déplacements très probables et pourtant très éloignés géographiquement, et inversement. La densité de déplacements effectués dans un voisinage fixe et pourtant peu probables et représentée sur la figure 2. Nous observons une majorité de POIs très peu accessibles, et une proportion non négligeable de POIs très accessibles.

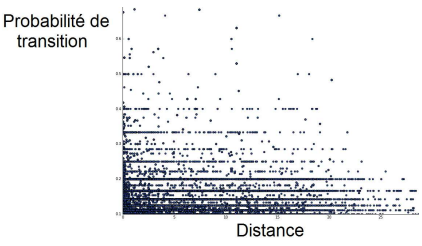

FIG. 1: Probabilité de transition en fonction de la distance associée.

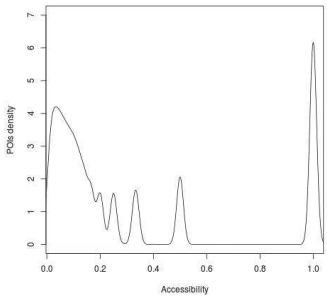

FIG. 2: Densité du nombre de POIs dans un voisinage fixe en fonction de leur accessibilité.

Résultats Après apprentissage des facteurs latents par variation d'inférence, nous avons comparé nos approches à la NMF (Lee et Seung (2001)), PMF (Salakhutdinov et Mnih (2007)) et la PF de base. La PF standard [3] a été définie dans l'équation 2. Pour

2. Le jeu de données est accessible à cette url : http ://www3.ntu.edu.sg/home/gaocong/datacode.htm
3. Une partie du code utilisé pour nos expérimentations est accessible ici : https ://github.com/ajbc/spf

mesurer la qualité des modèles que nous avons testés, nous avons utilisé la précision et le recall comme principales métriques parmi les nombreuses métriques alternatives existantes car ils sont largement utilisés dans les travaux connexes à notre approche. Nous observons que GeoSPF est le modèle qui donne les meilleurs résultats. Nous constatons une augmentation signicative de la qualité de la recommandation. En effet en se basant sur la mesure de recall@5, GeoPF améliore par rapport à la PF standard, le recall de 35% (et respectivement de 55% pour GeoSPF). Ce gain est significatif et confirme la validité de notre approche.

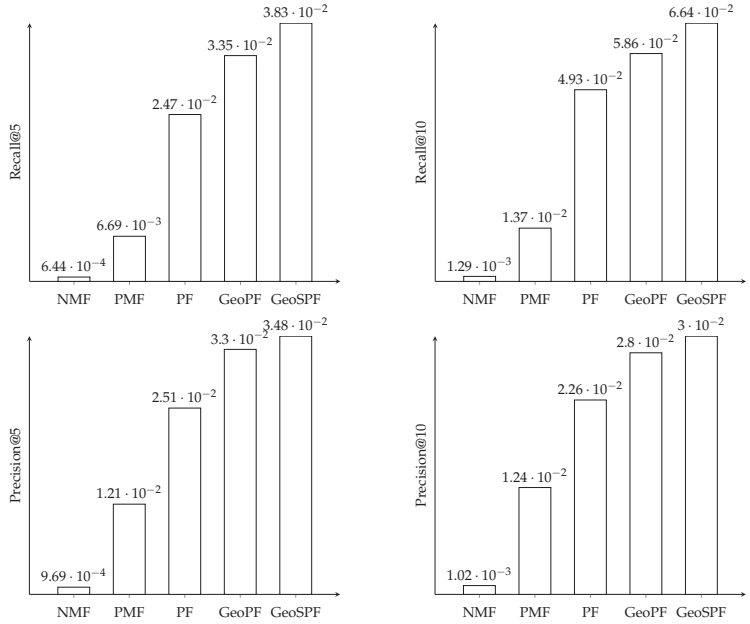

FIG. 3: Recall@K et Precision@K de différentes méthodes de factorisation sur le jeu de données de Foursquare.

5 Conclusion

La recommandation de POIs à partir des données issues des LBSNs comporte un certain nombre de spécificités qui rendent inefficaces les solutions de recommandation classiques. Dans ce premier travail sur les modèles de factorisation de Poisson, nous avons proposé une approche pour faire de la recommandation de POIs personnalisée qui puisse relever ces défis. Basés sur le concept d'accessibilité que nous avons proposé, nous avons réussi à obtenir des résultats prometteurs pour la suite. Une version plus complète de cet article est en particulier accessible librement [4]. Bien que notre approche fonctionne de façon satisfaisante sur des données issues des LBSNs, nous

4. Via cette url : http ://griesner.net/articles/egc17_article_long.pdf

pouvons observer qu'elle ne prend pas en compte le temps, ni davantage de variables latentes géographiques. Nous réservons donc l'intégration des influences temporelle et de la répartition des checkins dans des régions pour un travail futur.

6 Remerciements

Ce travail a été en partie financé par la Chaire de recherche de Télécom ParisTech sur le Big Data et la connaissance de marché. Pierre Dosne, ingénieur de recherche à Télécom ParisTech, nous a apporté une aide conséquente dans les expérimentations.

Références

Chaney, A. J., D. M. Blei, et T. Eliassi-Rad (2015). A probabilistic model for using social networks in personalized item recommendation. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, New York, NY, USA, pp. 43–50. ACM.

Cheng, C., H. Yang, I. King, et M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. AAAI, 2012.

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

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

Lee, D. D. et H. S. Seung (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, et V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13*, pp. 556–562. MIT Press.

Salakhutdinov, R. et A. Mnih (2007). Probabilistic matrix factorization. In J. C. Platt, D. Koller, Y. Singer, et S. T. Roweis (Eds.), *Advances in Neural Information Processing Systems 20, Canada, December 3-6, 2007*, pp. 1257–1264. Curran Associates, Inc.

Summary

The rapid growth of data volumes shared on location-based social networks (LBSN) enables the extraction of users' preferences. Then those preferences can be used to recommend to the user a list of points-of-interest matching his profile. Today the recommendation of points-of-interest has become an essential component of LBSN. Unfortunately traditional recommendation methods fail to adapt to the specific constraints of LBSN such as the high sparsity of the data, or to take into account the geographical influence. In this paper we present a model of recommendation based on the Poisson factorization that offers an effective solution to these constraints. We have tested our model through experiments on a realistic data set from the LBSN Foursquare. These experiences have enabled us to demonstrate a better recommendation than 3 models of state-of-the art.