L'exploitation de données contextuelles pour la recommandation d'hôtels

Marie Al Ghossein*,**, Talel Abdessalem*,***, Anthony Barré**

*LTCI, Télécom ParisTech, Université Paris-Saclay, 75013 Paris, France
**AccorHotels, Paris, France
***UMI CNRS IPAL, National University of Singapore
{marie.alghossein, talel.abdessalem}@telecom-paristech.fr, anthony.barre@live.fr

Résumé. Les systèmes de recommandation ont pour rôle d'aider les utilisateurs submergés par la quantité d'information à faire de bons choix à partir de vastes catalogues de produits. Le déploiement de ces systèmes dans l'industrie hôtelière est confronté à des contraintes spécifiques, limitant la performance des approches traditionnelles. Les systèmes de recommandation d'hôtels souffrent en particulier d'un problème de démarrage à froid continu à cause de la volatilité des préférences des voyageurs et du changement de comportements en fonction du contexte. Dans cet article, nous présentons le problème de recommandation d'hôtels ainsi que ses caractéristiques distinctives. Nous proposons de nouvelles méthodes contextuelles qui prennent en compte les dimensions géographique et temporelle ainsi que la raison du voyage, afin de générer les listes de recommandation. Nos expérimentations sur des jeux de données réels soulignent la contribution des données contextuelles à l'amélioration de la qualité de recommandation.

1 Introduction

Les systèmes de recommandation ont été principalement introduits pour aider les utilisateurs à faire face à la surcharge d'information sur le Web et pour augmenter le profit des entreprises. Leur tâche est de faire des suggestions d'items personnalisées à un utilisateur. Plusieurs domaines d'application ont bénéficié du développement de ces systèmes utilisés pour recommander des films à Netflix, des produits à Amazon, ou encore de la musique à Spotify. Néanmoins, le problème de recommandation d'hôtels se distingue des autres problèmes et soulève des difficultés particulières.

Les systèmes de recommandation d'hôtels souffrent d'un problème de démarrage à froid continu et cela pour trois raisons principales. Tout d'abord, le voyage est une activité relativement rare et la majorité des gens réservent un hôtel une ou deux fois par an. Ensuite, l'intérêt des voyageurs est susceptible de changer avec le temps. Ce phénomène est en particulier observé chez les personnes qui changent de statuts et passent de la réservation d'hôtels de la classe économique à la réservation d'hôtels luxueux. Enfin, la sélection d'un hôtel est généralement influencée par plusieurs facteurs contextuels comme par exemple la localisation, la tempora-

lité, la météo et la raison du voyage. Les systèmes de recommandation basés sur le contexte offrent une solution efficace pour faire face à ces difficultés (Adomavicius et Tuzhilin (2011)).

Dans cet article, nous proposons un système de recommandation d'hôtels qui combine deux nouvelles approches prenant en compte les dimensions géographique et temporelle ainsi que la raison du voyage. Le système regroupe dans un premier temps les utilisateurs qui partagent les mêmes affinités concernant les destinations visitées ainsi que les périodes de visite. Les modèles de recommandation sont ensuite construits indépendamment pour chaque groupe d'utilisateurs et l'apprentissage des modèles est guidé par la raison du voyage. Nos expérimentations sur des jeux de données réels extraits des bases de données d'AccorHotels [1] démontrent la contribution des données contextuelles à améliorer la qualité de recommandation.

La suite de cet article s'organise comme suit. Dans la section 2, nous donnons un aperçu de quelques approches de recommandation. Les sections 3 et 4 décrivent les méthodes sur lesquelles reposent le système de recommandation proposé. Enfin, nous présentons nos résultats expérimentaux dans la section 5, avant de conclure dans la section 6.

2 Préliminaires

Les systèmes de recommandation gèrent trois types d'entités : les utilisateurs, les items à recommander et les interactions entre utilisateurs et items. On distingue principalement deux approches pour la recommandation : le filtrage basé sur le contenu et le filtrage collaboratif. Dans le filtrage basé sur le contenu, les attributs descriptifs des items sont utilisés pour recommander à l'utilisateur des items similaires à ceux qu'il avait appréciés dans le passé. Les techniques de filtrage collaboratif exploitent la mesure de similarité entre les ensembles d'interactions effectuées par chaque utilisateur afin de sélectionner les items à recommander.

Pour traiter le problème de démarrage à froid continu rencontré lors de la recommandation d'hôtels, nous avons recours aux systèmes de recommandation basés sur le contexte. Ces systèmes tentent de reproduire le processus de prise de décision des voyageurs en considérant les facteurs contextuels qui influencent en général leurs décisions. Les recommandations sont alors guidées par le contexte actuel de l'utilisateur et par le comportement passé d'autres utilisateurs dans des situations similaires. Construire un système de recommandation robuste consiste alors à identifier les facteurs contextuels qui impactent les utilisateurs et à développer des modèles qui prennent en compte ces facteurs.

Cependant, les approches préalablement proposées échouent à s'adapter aux contraintes propres au problème de recommandation d'hôtels. La grande majorité des approches est conçue de façon à manipuler des données collectées explicitement comme les notes ou les commentaires. Elles ne peuvent être facilement transposées pour gérer des données collectées implicitement comme les réservations d'hôtels (utilisées dans notre cas de figure) ou les données de navigation. Par ailleurs, le contexte est traditionnellement modélisé comme une variable multidimensionnelle où chaque dimension correspond à un facteur contextuel. Les approches proposées supposent que les facteurs contextuels contribuent de façon égale au processus de prise de décision de l'utilisateur. Sachant que les voyageurs donnent plus de priorité à certains facteurs plutôt qu'à d'autres, ces approches ne sont pas adaptées au problème considéré.

1. http ://www.accorhotels.com

TAB. 1 – *Exemples de clusters de pays de résidence (liste non-exhaustive par cluster).*

Cluster 1	France, Belgique, Andorre, Guinée, Guyane, Algérie, Tunisie, Sénégal
Cluster 2	Italie, Allemagne, Suisse, Pays-Bas, Grèce, Roumanie, Russie, Turquie
Cluster 3	États-Unis, Canada, Pérou, Portugal, Inde, Chine, Corée du Sud, Vietnam
Cluster 4	ÉAU, Bahreïn, Koweït, Qatar, Maroc, Espagne, Tanzanie, Cuba

Une étude menée avec des experts du domaine révèle que les voyageurs sont influencés par plusieurs types de contexte qui sont principalement le contexte physique, social et psychologique. Le système de recommandation présenté intègre les dimensions géographique et temporelle (contexte physique) et la raison du voyage (contexte psychologique).

3 Intégration des influences géographique et temporelle

Nos jeux de données contiennent des utilisateurs répartis partout dans le monde et des hôtels répandus dans plus de 90 pays. La sélection d'un hôtel est généralement conditionnée par la destination à visiter. Nous considérons deux facteurs influençant le choix de destinations. Tout d'abord, le pays de résidence de l'utilisateur joue un rôle important dans le processus de décision. Les résidents d'un même pays suivent en général les mêmes tendances et visent dans la majorité de leurs déplacements des régions relativement proches du lieu de résidence. Ensuite, la popularité des destinations change en fonction des périodes de l'année marquée par les vacances et les saisons.

Afin d'intégrer l'influence des dimensions géographique et temporelle sur les utilisateurs, nous proposons de regrouper les pays de résidence dont les résidents ont des comportements similaires à l'égard du choix des destinations et des périodes de visite. La répartition des utilisateurs dans des clusters découle de ce regroupement de pays. Les modèles de recommandation sont ensuite construits indépendamment pour chaque cluster d'utilisateurs. Ces modèles, dits locaux (par opposition au modèle global construit pour l'ensemble des utilisateurs), sont capables de capter des préférences de granularité plus fine puisqu'ils couvrent des utilisateurs qui partagent des goûts similaires vis-à-vis des destinations.

Un pays de résidence est représenté par un vecteur comprenant un élément par destination et par mois. La valeur de cet élément est égale à la proportion de résidents ayant visité la destination en question durant le mois concerné. On applique l'algorithme K-moyennes afin d'obtenir K clusters de pays de résidence. La table 1 montre des exemples de pays par cluster.

Le clustering des pays de résidence suppose que les utilisateurs dont on dispose constituent un échantillon représentatif de la population de chaque pays. Le clustering dépend aussi de la distribution des hôtels AccorHotels dans tous les pays, sachant que celle-ci n'est pas uniforme.

4 Intégration de la raison du voyage

Le comportement du voyageur n'est pas le même lors d'un voyage d'affaires ou d'un voyage pour le plaisir. La raison du voyage n'est pas fournie explicitement par l'utilisateur mais peut être inférée à partir d'autres caractéristiques du séjour. Celles-ci couvrent, entre autres, le délai entre la date de réservation et la date du séjour, les jours de la semaine pendant lesquels le

séjour est effectué (week-end ou non) et le nombre d'adultes et/ou d'enfants. D'autre part, on note que les préférences des utilisateurs peuvent changer entre deux réservations successives.

Nous proposons un modèle qui prend en compte la raison du voyage afin d'améliorer l'apprentissage des paramètres du modèle de recommandation. Le modèle proposé donne aussi plus de poids aux interactions les plus récentes qui représentent mieux les préférences actuelles de l'utilisateur. Le modèle se base sur BPR (Rendle et al. (2009)), un modèle de factorisation de matrices, où nous utilisons un échantillonnage non uniforme du jeu de données afin d'apprendre le modèle.

Soit U l'ensemble des utilisateurs de taille m et H l'ensemble des hôtels de taille n. Soit $R \in \mathbb{R}^{m \times n}$ la matrice utilisateur-hôtel contenant m utilisateurs et n hôtels. La valeur de r_{uh} est égale à 1 si l'utilisateur u a visité l'hôtel h, et 0 sinon. Le but de la factorisation de matrices est d'approximer la matrice R par le produit de deux matrices de facteurs latents $P \in \mathbb{R}^{m \times k}$ et $Q \in \mathbb{R}^{k \times n}$, sachant que $k \ll \min(m, n)$.

Plusieurs méthodes ont été proposées afin d'apprendre les paramètres du modèle, i.e., les matrices P et Q. BPR (Rendle et al. (2009)) considère des paires d'items lors de l'apprentissage du modèle et optimise le ranking de ces items l'un par rapport à l'autre. L'hypothèse principale est qu'un item h observé pour un utilisateur u est préféré par rapport à un item h' non observé pour u. Les données considérées pendant la phase d'apprentissage sont représentées par l'ensemble suivant :

$$D_S := \{(u, h, h') \mid r_{uh} = 1 \wedge r_{uh'} = 0\}$$

où les triplets sont échantillonnés uniformément du jeu de données à cause du grand nombre de paires (h, h'). Nous ajoutons deux hypothèses au modèle existant.

Premièrement, nous supposons qu'un hôtel réservé par un utilisateur dans le cadre d'un voyage est préféré par rapport aux hôtels généralement choisis lors de voyages du même type. Soit I_{uh} l'ensemble des caractéristiques du séjour (citées ci-dessus) qui définissent le type du voyage. Nous proposons d'échantillonner les triplets de D_S et de D_S^{int} alternativement. D_S^{int} est défini de façon similaire que D_S et la probabilité d'échantillonner h' est la suivante :

$$p(h' \mid I_{uh}) \propto |\{r_{ah'} = 1 \mid I_{ah'} = I_{uh}, \forall a \in U\}|$$

La demande sur certains hôtels est plus forte lors des voyages dans certains contextes (voyage d'affaires ou pour le loisir). Ces hôtels ont alors plus de chance d'être sélectionnés comme item négatif pour ce contexte particulier.

Deuxièmement, nous supposons que les hôtels réservés récemment sont préférés par rapport aux choix plus anciens. Nous proposons alors d'échantillonner, en plus, d'un autre ensemble D_S^{rec}, défini comme suit :

$$D_S^{rec} := \{(u, h, h') \mid r_{uh} = 1 \wedge r_{uh'} = 1 \wedge rec(u, h, h') = 1\}$$

où $rec(u, h, h')$ est égal à 1 si la réservation faite par u dans h est plus récente que celle faite dans h', et 0 sinon.

Les trois ensembles D_S, D_S^{int} et D_S^{rec} sont utilisés pour apprendre le modèle. Nous introduisons les paramètres γ_{int} et γ_{rec} indiquant les probabilités de tirer les triplets de chacun des trois ensembles.

5 Évaluation expérimentale

Dans cette section, nous présentons une validation expérimentale des deux parties du système de recommandation proposé en utilisant un jeu de données réel issu des bases de données

TAB. 2 – $recall@N$ *et* $NDCG@N$ *pour Knni et BPR sous globalRS et localRS_K.*

Méthode	Métrique	globalRS	localRS_5	localRS_10	localRS_15
Knni	**recall@5**	0,0846	0,0861	0,08590	**0,0863**
	recall@10	0,1284	**0,1306**	0,1301	**0,1306**
	NDCG@5	0,0667	**0,069**	0,0689	**0,069**
	NDCG@10	0,0823	0,0848	0,0847	**0,085**
BPR	**recall@5**	0,3212	0,3253	**0,3261**	0,3258
	recall@10	0,3704	0,3740	**0,3741**	**0,3741**
	NDCG@5	0,2873	0,302	**0,303**	0,3028
	NDCG@10	0,3048	0,3194	**0,3201**	**0,3201**

d'AccorHotels.

Nous sélectionnons les utilisateurs ayant fait au moins une réservation depuis 4 ans. Le jeu de données utilisé contient 7,8 millions utilisateurs, 4,5 mille hôtels et 34 millions réservations. Nous divisons notre jeu de données en un ensemble d'apprentissage et un ensemble de test. Pour chaque utilisateur nous incluons 20% des hôtels visités les plus récents dans l'ensemble de test, et les 80% restants constituent l'ensemble d'apprentissage. Nous évaluons le $recall@N$ et le $NDCG@N$ du système (Shani et Gunawardana (2011)).

Afin d'évaluer la première partie du système, nous comparons les deux approches suivantes. Dans *globalRS*, un seul modèle de recommandation est construit pour tous les utilisateurs. Dans *localRS_K*, un modèle de recommandation est construit par cluster d'utilisateurs, pour un total de K clusters. Les résultats sont affichés dans la table 2 et concernent les deux modèles de recommandation suivants : Knni (Sarwar et al. (2001)) basé sur la recherche de plus proches voisins et BPR (Rendle et al. (2009)). La qualité de recommandation est améliorée sous *localRS_K* pour tout K. Après avoir mené un grand nombre d'expériences, nous fixons K à 10, maximisant le gain obtenu sur un grand nombre de modèles par rapport à *globalRS*.

Ensuite, nous comparons les méthodes suivantes :

— **BPR** (Rendle et al. (2009)) avec $k = 100$ et $\lambda_* = 0,0025$.
— **BPR$_{\text{int}}$** utilise les ensembles D_S et D_S^{int} avec $k = 100$, $\gamma_{int} = 0,5$ et $\lambda_* = 0,0025$.
— **BPR$_{\text{rec}}$** utilise les ensembles D_S et D_S^{rec} avec $k = 100$, $\gamma_{rec} = 0,1$ et $\lambda_* = 0,0025$.
— **BPRx3** utilise les ensembles D_S, D_S^{int} et D_S^{rec} avec $k = 100$, $\gamma_{rec} = 0,1$, $\gamma_{int} = 0,6$ et $\lambda_* = 0,0025$.

Nous reportons dans la table 3 les résultats pour un cluster d'utilisateurs qui en comprend environ 60 mille, sachant que les conclusions sont similaires pour les autres clusters. Les résultats montrent l'importance de l'intégration des données liées à la raison du voyage et de la mise en valeur des interactions les plus récentes. Ceci a un impact sur l'apprentissage du modèle qui génère de meilleures recommandations.

6 Conclusion

Nous proposons dans cet article un système de recommandation d'hôtels basé sur le contexte afin de faire face au problème de démarrage à froid continu rencontré lors de la re-

TAB. 3 – *$recall@N$ et $NDCG@N$ pour les variantes de BPR.*

Métrique	BPR	BPR_{rec}	BPR_{int}	BPRx3
recall@5	0,3396	0,3703	0,3497	**0,3734**
recall@10	0,4328	0,4457	0,4376	**0,4477**
NDCG@5	0,3146	0,3548	0,3245	**0,3577**
NDCG@10	0,351	0,3847	0,3588	**0,3871**

commandation d'hôtels. Nous intégrons dans un premier temps les dimensions géographique et temporelle et regroupons ensemble les utilisateurs soumis aux mêmes influences. Nous construisons ensuite un modèle de recommandation par groupe d'utilisateurs et utilisons les caractéristiques liées à la raison du voyage afin d'améliorer l'apprentissage du modèle. Nous envisageons l'intégration d'autres facteurs contextuels dans de futurs travaux, comme la météo ou les points d'intérêt.

Références

Adomavicius, G. et A. Tuzhilin (2011). Context-aware recommender systems. In *Recommender systems handbook*, pp. 217–253. Springer.

Rendle, S., C. Freudenthaler, Z. Gantner, et L. Schmidt-Thieme (2009). Bpr : Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pp. 452–461. AUAI Press.

Sarwar, B., G. Karypis, J. Konstan, et J. Riedl (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pp. 285–295. ACM.

Shani, G. et A. Gunawardana (2011). Evaluating recommendation systems. *Recommender systems handbook*, 257–297.

Summary

In recent years, recommender systems have witnessed an increased interest from industry and academia. The deployment of such systems in the hotel industry needs to satisfy specific constraints, making the direct application of classical approaches insufficient. There is an inherent complexity to the problem, starting from the decision-making process for selecting accommodations, which is sharply different from the one for acquiring tangible goods, to the multifaceted behavior of travelers, often selecting accommodations based on contextual factors. Travelers recurrently fall into the cold-start status due to the volatility of interests and the change in attitudes depending on the context. In this paper, we propose a context-aware recommender system for hotel recommendation. The system is based on two novel approaches that take into account geography, temporality, and the trips' intent. Our experiments on a real-world dataset show the impact of taking into account contextual data in improving the quality of the recommendation.